IJACSA
COMPUTER SCIENCE JOURNAL
International Journal of Advanced Computer Science and Applications

Volume 17 Issue 6
June 2026

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

THE SCIENCE AND INFORMATION ORGANIZATION

www.ijacsa.thesai.org

Table of Contents

Paper 1: A Multi-Stage Framework for Bias Detection and Mitigation in AI-Driven Recruitment Systems

Authors: Gideon Assafuah, Claude Turner, Carlene Turner, Kingsley Nwosu

PAGES 1 - 16

Paper 2: Zero-Disclosure Material Passports for Verifiable Provenance in Multi-Tier Supply Networks

Authors: Shivani Dharmavaram

PAGES 17 - 24

Paper 3: Exploring the Structure Resulting from Unstructured Neural Network Pruning

Authors: Jamil Gafur, Max Milkert, Kevin Patrick Griffin, Nicholas T. Wimer, Charles Tripp, Steve Goddard

PAGES 25 - 34

Paper 4: Privacy Leakage and Memorization in Fine-Tuned Clinical Language Models: A Controlled Study of Defenses and Backbone Choice on Clinical Narrative Transcriptions

Authors: Yassine Chahid, Anas Chahid, Ismail Chahid, Aissa Kerkour Elmiad

PAGES 35 - 44

Paper 5: Deployer-Side XAI Instrumentation for Regulated AI: A Clinical Case Study in ICL Sizing

Authors: Dorleta Urrutia-Onate, Enrique Onieva, Asier Perallos

PAGES 45 - 61

Paper 6: DcDM: A Pre-training Data Evaluation Framework for Proactive Drift Prevention in Machine Learning

Authors: Okjoo Choi, Wonsun Shin

PAGES 62 - 72

Paper 7: Characterizing Operational Drift in Cement Manufacturing Process Data via Self-Supervised Representation Learning

Authors: Changgyun Kim

PAGES 73 - 81

Paper 8: Digital-to-Physical Transfer of Adversarial Patches for Aerial Vehicle Detection

Authors: Jung Heum Woo, Eun-Kyu Lee

PAGES 82 - 93

Paper 9: A Hybrid Ethereum-Based Architecture for Secure Electronic Health Records: Consent, Integrity Anchoring and Auditable Access

Authors: Rodica Doina Zmaranda, Attila-Imre Kovacs, Daniela Elena Popescu, Alexandrina Mirela Pater

PAGES 94 - 103

Paper 10: Using Artificial Intelligence Techniques for Error Detection and Repair in Software Development

Authors: Abdulaziz Aladwani, Sultan Alsamaani, Turki Alrumaykhani, Mohamed Tahar Ben Othman

PAGES 104 - 113

Paper 11: Analyzing User Experience in Mobile Banking Applications Through Text Mining

Authors: Bora Lamaj (Myrto), Markela Muça, Klodiana Bani

PAGES 114 - 119

Paper 12: Design and Evaluation of a 13K Ultra-High-Resolution Web-Based Virtual Reality Platform for Immersive Spatial Visualization

Authors: Worapon Manosroi, Apisak Phromfaiy, Pitak Khlaichom

PAGES 120 - 126

Paper 13: Artificial Intelligence in Software System Quality Assurance: A Systematic Literature Review of Techniques, Tools, and Challenges (2016–2026)

Authors: Abdullah A H Alzahrani

PAGES 127 - 143

Paper 14: A Comparative Evaluation of Large Language Models for Named Entity Recognition in Cyber Threat Intelligence

Authors: Aykhan Huseynli

PAGES 144 - 153

Paper 15: A QR-Code-Based Mobile Learning System for Science Instruction in Resource-Limited Schools

Authors: Gina B. Selga

PAGES 154 - 165

Paper 16: An Improved Pre-processing Method for High-Quality MRI Images in Brain Tumor Detection

Authors: Nirmala, Kavitha B C

PAGES 166 - 174

Paper 17: Vibration-Aided Picture-Based Authentication for Shoulder-Surfing Resistant Mobile Login

Authors: Ibrahim Albadi, Mahdi Almalki, Abdullah Albokhari, Salman Alsumairi, Sami Atiyah, Faisal Alsubaei, Abdullah Abuhussein

PAGES 175 - 187

Paper 18: AI-Based Career Transition Recommendation System Using Controlled Educational Data Augmentation

Authors: Gerlix ADANKON, Pelagie HOUNGUE, Melckior DEGBOE, Corelle GOGAN

PAGES 188 - 201

Paper 19: Verifiable Learned PR-Tree Indexing for Privacy-Preserving Range Queries Over Encrypted Geospatial Data

Authors: Anagha Aher, Sangita Chaudhari

PAGES 202 - 210

Paper 20: A Multi-Modal Deep Learning Framework for Student Soft Skills Development in Adaptive Learning Environments

Authors: Marzhan Bekbolat, Kamalbek Berkimbayev, Rustam Abdrakhmanov, Serik Kenesbayev, Nuraim Ibragimova, Zhaksylyk Dzhanabayev

PAGES 211 - 226

Paper 21: A Hybrid Multi-Objective AI Framework for Curriculum-Aware Examination Generation

Authors: Mohamed Fathy Yehia, Yehia M. Helmi, Mahmoud Mohamed Bahloul

PAGES 227 - 236

Paper 22: An Intelligent Scheduling Optimization Algorithm for Multimodal Cache Resources with Status Awareness of Metropolitan Area Network CDN Nodes

Authors: Ruirong Jiang, Zhibiao Xiong, Junliang Wu, Jinyong Xu

PAGES 237 - 246

Paper 23: MSE-Guided Hybrid U-Net Framework for Automatic Kidney Segmentation and Spatial Localization

Authors: Dannial Asyraf Shahrul Anuar, Nabilah Ibrahim, Audrey Huong

PAGES 247 - 254

Paper 24: Learning Mathematics Through Play: Design and Development of a Serious Game for Undergraduate Students

Authors: Nur Nabilah Abdul Razak, Nur Nabila Farhana Mohd Noh, Nur Dayana Huda Mohd Zol Azhari, Ratna Zuarni Ramli, Ahmed M S Elaklouk, Azyan Yusra Kapi

PAGES 255 - 261

Paper 25: Applying the AuRa Consensus Model for Digital Certificate Management in a Private Ethereum Blockchain

Authors: Robiah Arifin, Wan Aezwani Wan Abu Bakar, Mustafa Man, Mohamad Afendee Mohamed, Evizal Abdul Kadir

PAGES 262 - 270

Paper 26: A Bibliometric Analysis of Internet of Things and Augmented Reality Applications in Agriculture: A Scopus-Based Review (2019–2026)

Authors: Ruziana Mohamad Rasli, Sobihatun Nur Abdul Salam

PAGES 271 - 280

Paper 27: Evolutionary Allocation with Dynamic Guidance for Pre-Scheduled Timetables

Authors: Anniza Hamdan, Sze San Nah, Goh Say Leng, Emily Sing Kiang Siew

PAGES 281 - 288

Paper 28: Multi-Class Stress Detection Using Electrodermal Activity: Evaluation of A Hybrid Model Using UBFC-Phys Dataset

Authors: Kawther Alsayed, Hamza Ghandorh, Wael M.S. Yafooz

PAGES 289 - 297

Paper 29: ATC Automation Based on Dynamic Aircraft Separation Techniques Accounting for Weather Conditions

Authors: Omayma Raziq, Mohamed El Khaili, Hasna Nhaila, Azeddine Khiat

PAGES 298 - 305

Paper 30: A Computational Analysis of Housing Affordability: Robust Bootstrap Regression Modeling of the Safety Premium

Authors: Natia Terterashvili, Shota Shaburishvili

PAGES 306 - 311

Paper 31: An Enhanced Kalman Filter-Based Hybrid Battery Management System for Energy Optimization in IoT-Enabled Smart Agriculture

Authors: Mohd Kamir Yusof, Nur Yasmin Salleh, Senny Luckyardi,
Wan Mohd Amir Fazamin Wan Hamzah, Mustafa Man

PAGES 312 - 319

Paper 32: Enhanced Real-Time Fire Detection Systems Using Deep Learning and Differentiating Between Dangerous and Non-Dangerous Fires

Authors: Mohamed Youssef, Mohamed Marie, Sarah Naiem

PAGES 320 - 329

Paper 33: A Novel Hybrid Deep Learning Validation Model for Real-Time and Synthetic Image Inputs in Capsicum Plant Disease Diagnosis

Authors: Prashant Vikhe, Baisa Gunjal

PAGES 330 - 336

Paper 34: Large-Scale Static Malware Detection Using Classical Machine Learning Models: An Evaluation on the EMBER Dataset

Authors: Achmad Fauzan, Tito Pinandita, Aulia Desy Nur Utomo

PAGES 337 - 347

Paper 35: A Data-Driven Visual Analytics Framework for Transaction-Level Retail Profit Modeling and Decision Support

Authors: Donia Badawood

PAGES 348 - 362

Paper 36: Fostering Programming Interest Through Block Coding in Young Learners

Authors: Dorjpalam Tserendejid, Khajidmaa Otgonbaatar, Magsarjav Bataa

PAGES 363 - 370

Paper 37: MYPO-Net: A Robust Deep Learning Approach for Multi-Yoga Pose Detection and Occlusion Handling

Authors: Rehana Danial, Nosheen Qamar, Nosheen Sabahat, Faria Nazir, Ali Salem Bin Sama, Lamia Hassan Rahamatalla, Osman Elwasila, Abdulrahman Alojail, Marwan Abu-Zanona

PAGES 371 - 382

Paper 38: Confirmatory Factor Analysis of Phishing Susceptibility in Indonesia

Authors: Harris Simaremare, Muhammad Fikri, Zarina Shukur

PAGES 383 - 394

Paper 39: AgricultureEarlyWarning: A Web-Based Climate Advisory and Early Warning Platform for Albanian Farmers

Authors: Irida Gjermeni

PAGES 395 - 403

Paper 40: Spatial Anemia and Osmotic Fragility of Erythrocytes in Microgravity Conditions: A Systematic Review

Authors: Lilian Ocares-Cunyarachi, Natalia Vargas-Cuentas, Avid Roman-Gonzales, Georgina Chávez, Brigitte Torres-Pinedo, Ana Huamani-Huaracca

PAGES 404 - 418

Paper 41: Artificial Intelligence-Assisted Community Needs Assessment and Extension Planning: Evidence from Wesleyan University-Philippines Partner Communities

Authors: Eufemia Ayro, Karl Leugim Bernarte, Hazel May Babiera, Evangeline Agpoon, Jhon Carlo Villa, Maureen Bondoc, Jennyfer Villalon, Jose Arsenio Adriano, Christian Navarro

PAGES 419 - 428

Paper 42: A Comparative Analysis of Decision Tree, Random Forest, and Logistic Regression Models in Predicting Business Readiness for Digital Technology Integration

Authors: Gloria M. Ducut

PAGES 429 - 437

Paper 43: Canthus-Scaled 468-Landmark FaceMesh Framework for Pupillary Distance Estimation Using Nested AutoML Calibration

Authors: Mohd Izzuddin Mohd Tamrin, Sherzod Turaev, Takumi Sase, Mohd Zulfaezal Che Azemin, Tengku Mohd Tengku Sembok

PAGES 438 - 445

Paper 44: A Hybrid CNN-Ensemble Framework for Robust DeepFake Image Detection

Authors: Mohammad Alsulami

PAGES 446 - 457

Paper 45: A Bibliometric Mapping of Transformer-Based Misinformation Detection: Trends and Gaps

Authors: Borhan Ab Rahman, Mohd Zakree Ahmad Nazri, Mohd Ridzwan Yaakub

PAGES 458 - 468

Paper 46: Web System to Optimize Information Management for Beca 18 Selection-Stage Applicants, 2025

Authors: Jack Quispe-Vivas, Cristian Sedano-Contreras, Junior Segovia-Chalco, Brian Meneses-Claudio

PAGES 469 - 474

Paper 47: A Modified Lyapunov-Based MEC Offloading Algorithm for Severity-Aware QoS in 5G/B5G IoT Systems

Authors: Sahana S Reddy, R. Sukumar

PAGES 475 - 489

Paper 48: Bridging Topic Modelling Outputs to Bayesian Hierarchical Model Using LLM and WordNet Parameter Labelling

Authors: Vadrianey Asas, Sarah Samson Juan, Stephanie Chua, Evan Lau, Jane Labadin

PAGES 490 - 500

Paper 49: Artificial Intelligence and the Transformation of Academic Integrity in Higher Education: A Systematic Review

Authors: Wannakorn Phornprasert, Wongpanya S. Nuankaew, Pratya Nuankaew

PAGES 501 - 511

Paper 50: A Secure Integrated Cloud Storage Framework Using Lorenz Chaotic Key Generation, AES-256-GCM, and PBFT-Based Blockchain Verification

Authors: Walde Rajesh Baliram, Bashir Alam, Mohammad Najmud Doja

PAGES 512 - 531

Paper 51: Comparative Evaluation of Traditional and Transformer-Based Models for Risk-Level Classification of Uzbek Telegram Messages

Authors: Feruzakhon A. Qoyliyeva, Ozod J.Babomuradov, Akmal A. Savurbayev

PAGES 532 - 536

Paper 52: MedChain: A Privacy-Preserving Framework for Scalable Electronic Health Record Sharing on Blockchain and InterPlanetary File System

Authors: S. Venkateswaran, N. Vijayaraj

PAGES 537 - 548

Paper 53: Token-Level PII Detection with Symbolic, Sequential, and Transformer-Based Ensemble Models

Authors: Hessah Abdullah Alshamrani, Mona Alnahari

PAGES 549 - 562

Paper 54: Holistic Model of Advanced Analytics to Optimize Organizational Decision-Making

Authors: Juan Carlos Morales-Arevalo, Ciro Rodríguez

PAGES 563 - 575

Paper 55: Multi-Scale Curvelet-Based Directional Denoising for Chest X-Ray Images

Authors: Neenu Sebastian, B. Ankayarkanni

PAGES 576 - 589

Paper 56: Characterizing Failure Points in Rule-Based Spreadsheet Data Transformation: A Stage-Oriented Taxonomy and Empirical Failure Matrix

Authors: Fakhrul Adli Mohd Zaki, Mustafa Man, Mohamad Nor Hassan

PAGES 590 - 598

Paper 57: Arabic Sign Language Alphabet Recognition Using Transfer Learning: Evaluation, Ablation, and Deployment

Authors: Abdelfatah Maarouf, Otman Maarouf, Abdelaali Benaiss, Rachid El Ayachi, Mohamed Biniz

PAGES 599 - 607

Paper 58: Fish Disease Detection Using Modified Haar Wavelet with Adaptive Coefficient Selection

Authors: Tri Handayani, Nor Hazlyna Binti Harun

PAGES 608 - 617

Paper 59: Adaptive Neuro-Digital Twin with Cross-Domain Multimodal Representation Learning for Early Alzheimer's Disease Prognosis

Authors: V S Krushnasamy, Annapurna Mishra, Pratik Gite, Ganesh Kumar Anbazhagan, Adapa Gopi, M. Misba, A. Arul Anitha, Osama R. Shahin

PAGES 618 - 632

Paper 60: Semantic Style Transfer for Paintings Using Convolutional Neural Networks (CNNs)

Authors: Hafiz Muhammad Jamsheed Nazir, Zheng Jiangbin, Omar Alsaleh

PAGES 633 - 643

Paper 61: Multi-Dimensional Fractal Vulnerability Study Of Alzheimer's Brain Networks

Authors: Tin Tin Ting, Neha Hema Raj, Geetha N K, Thangaraj C, Olusegun D. Samuel

PAGES 644 - 650

Paper 62: Progression-Aware Temporal Graph Transformer for Reliable Chronic Kidney Disease Trajectory Prediction

Authors: Roshan D. Suvaris, Padmavathy E, Dilfuza Akabirkhodjaeva, T. K. Rama Krishna Rao, R. Sindhu, Farrukh Sobia, Elangovan Muniyandy, Aseel Smerat

PAGES 651 - 663

Paper 63: Toward Adaptive Educational Intervention: Meta-Adaptive Cross-Modal Gating for Few-Shot Personalized Intervention

Authors: Houda Kaa, Hanane Allioui, Ilham Oumaira

PAGES 664 - 677

Paper 64: An Adaptive Smart Business Intelligence Model Based on Enhanced HGO Discovery for Real-Time in Inpatient Care

Authors: Hengki, Rahmat Gernowo, Oky Dwi Nurhayati

PAGES 678 - 683

Paper 65: Multi-Relation Knowledge Graph-Guided Transformer with Objective-Aware Route Selection for Vehicle Routing Problems with Time Windows

Authors: Somkiat Kosolsombat, Chiabwoot Ratanavilisagul

PAGES 684 - 690

Paper 66: Deployment-Aware 30-Day Readmission Prediction in Resource-Limited Hospitals: Calibration, Threshold Policy, and Decision Utility

Authors: Samer Asad Malalha, Ma Burhanuddin, Hatem T M Duhair, Jamil Abedalrahim Jamil Alsayaydeh, Mazen Farid

PAGES 691 - 705

Paper 67: Expert Evaluation of the Proposed ERP Model for Legal Compliance and Adaptation

Authors: Shkëlqim Miftari, Azir Aliu, Artan Luma

PAGES 706 - 715

Paper 68: AI-Driven Service Innovation, Customer Satisfaction, and Guest Loyalty: Evidence from Shenyang Airport Hotel, China

Authors: Dayong Zu, Kawalin Angkananon, Yoksamon Jeaheng

PAGES 716 - 727

Paper 69: Ontology-Based Business Process Modeling: A Review

Authors: Low Kok Thai, Furkh Zeshan, Nazri Kama, Riza Sulaiman, Mohammad Nazir Ahmad, Uwais Qidwai

PAGES 728 - 745

Paper 70: A Robust Security Framework for Cloud Data Storage Using Lightweight Blockchain Technology

Authors: Renuka GOLLA BALA, S. Gnanavel

PAGES 746 - 756

Paper 71: ERP Implementation Success Factors Across Project Phases: An Action Research and Fuzzy AHP Study in Moroccan SMEs

Authors: Yassine Zouhair, Younous Elmrini

PAGES 757 - 769

Paper 72: Hybrid Data Fusion and Deep Learning for Dynamic Risk Index Modeling in Secure Learning Management Systems

Authors: Vani T, S. Sathya

PAGES 770 - 783

Paper 73: Adaptive Honey Badger Optimization with Bernoulli Chaotic Mapping and Decreasing Neighborhood for Efficient Cloud Task Scheduling

Authors: Yanfang XING

PAGES 784 - 792

Paper 74: Structure-Aware Latent Diffusion for High-Quality Line Art Colorization

Authors: Shuhua Xu, Qiang Ai, An Zhao, Guan Yang, Bo Chen

PAGES 793 - 803

Paper 75: C-TriLoRA: Cross-Corpus Kazakh SER via Tri-Factor LoRA and CORAL

Authors: Bakdaulet Kynabay, Aimoldir Aldabergen, Shirali Kadyrov

PAGES 804 - 811

Paper 76: Query Recovery Attack Based on Multi-Source Leakage and Semantic Embedding in Searchable Symmetric Encryption

Authors: Xiaogang Yuan, Xinle Yang, Dezhi An

PAGES 812 - 829

Paper 77: Multiplicative Gate State Space Models with Skip-Net for High Accuracy COVID-19 Time-Series Prediction

Authors: Krung Sinapiromsaran, Supakit Sroynam

PAGES 830 - 836

Paper 78: Topology-Aware Fingerprint Representation Using Graph Convolutional Network for Robust Minutiae Refinements

Authors: Yoogesh A, Rama Prasath A

PAGES 837 - 847

Paper 79: Architecting a Low-Latency RAG System for Fast-Moving Consumer Goods (FMCG) Customer Support: A Case Study in Industrial Software Deployment

Authors: Meredita Susanty, Alghifari Rasyid Zola

PAGES 848 - 855

Paper 80: Adaptive Hybrid Intrusion Detection for Realistic Zero-Day Attacks in Cloud and Edge Environments

Authors: Nithin U, Ganeshayya Shidaganti, Sangeetha V, Vishwachetan D

PAGES 856 - 863

Paper 81: Towards Reliable Recognition of Concurrent Abnormal Patterns in Control Charts Using Multi-Label Deep Learning

Authors: Mohammed Modar, Abdelilah Ganmati, Omar El Farissi

PAGES 864 - 872

Paper 82: Deep Learning for the Classification of Kidney Diseases in Medical Images Using ResNet-50 and Grad-CAM

Authors: Laberiano Andrade-Arenas, Inooc Rubio Paucar, Cesar Yactayo-Arias

PAGES 873 - 886

Paper 83: Case-Based Reasoning Model for Predicting the Malaria Cases

Authors: Konan N'gatta Aimé Kouassi, Koffi Kouakou Ive Arsene, Gooré Bi Tra

PAGES 887 - 896

Paper 84: A Hybrid Semantic-Statistical Feature Fusion Framework for Bilingual Text Classification on Multilingual Big Data Corpora

Authors: Kavitha M, Purohit Shrinivasacharya, Y S Nijagunarya

PAGES 897 - 912

Paper 85: Security from Design, Bridging Model-Driven Architecture and DevSecOps Using Zynerator

Authors: Younes Zouani, Mohamed Lachgar, Youssef Harrati, Mohamed Hanine, Sulieman S. Alshuhri, Amal Alomran

PAGES 913 - 925

Paper 86: Mitigating Data Migration Risks in the Cloud via GA-Optimized Hybrid Cryptography Mechanisms

Authors: Anjali Dhaman, Ugrasen Suman

PAGES 926 - 933

Paper 87: An Explainable Hybrid AI Framework for Climate-Driven Environmental Health Risk Prediction in Agro-Ecosystems

Authors: Fatima-Zahra Alaoui, Laila El Jiani, Sanaa El Filali, Rachida Ait Abdelouahid, Zouheir Banou

PAGES 934 - 947

Paper 88: Behavior-Aware Access Control for IoT Networks Using Lightweight Machine Learning at the Gateway Level

Authors: Yaseen Alduwayl, Abdullah Alessa, Mounir Frikha

PAGES 948 - 967

Paper 89: A New Drilling Rate of Penetration Prediction Model by Particle Swarm Optimization and Gradient Boosting Regression

Authors: Faris Aiman Jamaluddin, Marina Yusoff, Diva Kurnianingtyas, Mohamad Taufik Mohd Salledud-din

PAGES 968 - 977

Paper 90: An Experimental Evaluation of Deep Learning Networks for Automated Breast Cancer Detection

Authors: Partha Chakraborty, Umme Aiman Jannat

PAGES 978 - 988

Paper 91: Optimization of Service Function Chain Placement in Cloud-Fog-Edge Networks

Authors: Chandrapal Singh Dangi, Sanjay Sharma

PAGES 989 - 1001

Paper 92: 2I-CSO: A Novel Intelligent and Interoperable Cat Swarm Optimizer Approach to Optimal Cluster Head Selection and Self-Termination Search

Authors: Oumaima Hassan, Mohammed Essaid Riffi

PAGES 1002 - 1015

Paper 93: Emotion-Aware Serendipity Recommendation from Textual Reviews: BERT-Based Emotional Clustering

Authors: Mariam Benayad, Ahmed Zellou

PAGES 1016 - 1028

Paper 94: QuantumGuard: A Post-Quantum Resilient Deception-Driven Framework for Proactive Threat Hunting and Cognitive Honeynet Orchestration in 6G-Enabled Cyber-Physical Systems

Authors: Daifallah Zaid Alotaibe

PAGES 1029 - 1043

Paper 95: Predicting Histological Progression in Primary Biliary Cirrhosis Using Advanced Machine Learning Techniques

Authors: Laberiano Andrade-Arenas, Cesar Yactayo-Arias, Inoc Rubio Paucar

PAGES 1044 - 1057

Paper 96: Deep Learning for Canonical Reconstruction of Deformable Objects from Depth Images in Robotic Vision and SLAM-Aware Perception

Authors: Fahd Alhamazani

PAGES 1058 - 1064

Paper 97: Accelerating Blockchain Consensus: A Parallel Mining Approach for High-Throughput, Low-Latency Networks

Authors: Sohail Jabbar

PAGES 1065 - 1081

Paper 98: Explainable Neural Network Prediction of Post-COVID-19 Depression via Monte Carlo Simulation

Authors: Siham AKIL, Sara SEKKATE, Abdellah ADIB

PAGES 1082 - 1094

Paper 99: Context-Aware Sentiment Analysis of E-Commerce Reviews Using a BERT-CNN-BiLSTM Hybrid Model

Authors: Mahmud Reza Mahim, Fahad Bin Z Islam, Md Saklain Mahmud, Md. Imtiyaz Hasan, Sifat Rahman Ahona

PAGES 1095 - 1105

Paper 100: Advanced and Classical Selection Methods in Genetic Algorithms: A Comprehensive Comparative Analysis

Authors: Husam S. Mashaqbeh, Putra Sumari, Hamza A. Mashagba, Mohammed Hashem Almourish, Azlan Abd Aziz, Lara A. Al-Mashagba, Wael Waheed Alqassas

PAGES 1106 - 1114

Paper 101: Combatting Phishing Attacks: Leveraging Machine Learning for Real-Time Detection in Penetration Testing

Authors: Ashwag Alotaibi, Mounir Frikha

PAGES 1115 - 1123

Paper 102: A Model for Refactoring Monolithic Applications to Microservices Using Domain-Driven Design: A Case Study on IoT Irrigation Systems

Authors: Munezero Immaculee Joselyne, Ngenzi Alexander, Hitimana Eric, Ipinnimo Oluwafemi

PAGES 1124 - 1136

Paper 103: A Multi-Criteria Decision-Making Model for ERP Selection in Moroccan SMEs Using Fuzzy AHP

Authors: Yassine Zouhair, Younous Elmrini

PAGES 1137 - 1148

A Multi-Stage Framework for Bias Detection and Mitigation in AI-Driven Recruitment Systems

Gideon Assafuah, Claude Turner, Carlene Turner, Kingsley Nwosu
Department of Computer Science, Norfolk State University, Virginia, USA
Department of Sociology, Norfolk State University, Virginia, USA

***Abstract*—The use of machine learning in recruitment has raised growing concerns about fairness, as automated hiring systems can generate unequal outcomes across demographic groups. These disparities are influenced not only by imbalanced data but also by the behavior of learning algorithms, making bias a multidimensional challenge that cannot be effectively addressed with single-stage solutions. This study introduces an integrated framework for bias detection and mitigation in AI-driven recruitment systems, combining interventions at the data, model, and decision levels within a unified evaluation pipeline. The framework is assessed using multiple classification models of varying complexity and evaluated with established fairness metrics. In addition, explainability techniques are employed using SHAP-based feature attribution to investigate hidden dependencies and assess the sensitivity of predictions to demographic attributes. Experimental results show that baseline models achieve strong predictive performance, with accuracy ranging from 0.807 to 0.816; however, fairness evaluation reveals substantial disparities, with Disparate Impact as low as 0.190 and Demographic Parity Difference exceeding 0.27 in some cases. After applying the proposed multi-stage mitigation approach, fairness metrics improve significantly: Disparate Impact meets or exceeds the legal threshold of 0.80 across all models, with reductions in Demographic Parity Difference of 70–85% and Equal Opportunity Difference of 47–75%, and demographic disparities are reduced to 0.029–0.053. These improvements are achieved with minimal performance trade-off, as overall accuracy decreases by at most 4.2 % points while ROC AUC remains unchanged. The findings demonstrate that bias in recruitment systems arises from the interplay between data and model dynamics and highlight the importance of coordinated mitigation strategies throughout the machine learning lifecycle. This work provides a practical, scalable approach to developing fair and transparent AI systems for hiring applications.**

***Keywords*—Bias detection; bias mitigation; AI in recruitment; algorithmic fairness; explainable AI; fairness metrics; machine learning; responsible AI**

I. Introduction

The rapid adoption of artificial intelligence (AI) in recruitment and hiring processes has transformed how organizations evaluate and select candidates [30]. Automated systems are increasingly used to screen resumes, rank applicants, and support decision-making in large-scale recruitment settings, offering improved efficiency and scalability [1], [2], [28]. However, as these systems become more influential in determining access to employment opportunities, concerns have emerged regarding their fairness and potential to produce unequal outcomes across demographic groups [14], [15].

A growing body of research has shown that machine learning models can inherit and amplify historical biases embedded in training data, leading to systematic disparities in hiring outcomes. These biases may arise from imbalanced datasets, proxy variables, or algorithmic decision boundaries, and can disproportionately affect groups defined by sensitive attributes such as gender, race, or age [12], [16]. As a result, AI-driven recruitment systems risk reinforcing existing social inequalities while operating under the perception of objectivity, raising concerns about transparency, accountability, and ethical responsibility [7], [8].

Existing approaches to addressing bias in machine learning systems typically focus on isolated stages of the pipeline. Data-level methods aim to rebalance or transform datasets to reduce representation bias, while model-level techniques incorporate fairness constraints during training. Post-processing approaches, on the other hand, adjust model outputs to improve fairness after predictions have been generated. Although these methods have demonstrated effectiveness in specific contexts, their isolated application often fails to capture the complex ways bias propagates throughout the entire machine learning lifecycle [9]. Consequently, there is a need for integrated approaches that address bias holistically rather than at a single stage.

In addition to mitigation strategies, explainability has emerged as a critical component in understanding and evaluating fairness in AI systems. Techniques such as SHAP-based feature attribution and local explanation methods provide insights into how models make decisions and whether sensitive attributes or their proxies influence outcomes. These methods are particularly important in high-stakes domains such as recruitment, where opaque decision-making can undermine trust and accountability [10], [11].

This study addresses these challenges by proposing a multi-stage framework for bias detection and mitigation in AI-driven recruitment systems. Unlike prior approaches that focus on a single level of intervention, the proposed framework integrates data-level auditing, model-level fairness-constrained training, and a novel post-processing threshold optimizer within a unified pipeline. Critically, the three stages are designed to operate in sequence — pre-processing corrects data-level imbalance, in-processing embeds fairness constraints into model training, and post-processing applies a group-specific

threshold search governed by a composite objective function that explicitly encodes the EEOC 80% legal threshold as a hard constraint. This end-to-end coordination across all three intervention levels, with each stage informing the next, is what distinguishes the proposed approach from existing single-stage or loosely coupled multi-technique fairness methods. The framework is evaluated using multiple machine learning models, including Logistic Regression, Random Forest, and Neural Networks, to analyze how bias manifests across varying levels of model complexity. Fairness is assessed using established metrics such as Disparate Impact, Demographic Parity Difference, and Equal Opportunity Difference. In contrast, explainability techniques are used to uncover hidden dependencies and assess decision sensitivity through SHAP-based feature attribution.

The contributions of this work are fourfold, and collectively distinguish it from prior art in three important ways. First, while pre-processing, in-processing, and post-processing fairness techniques each exist in isolation in the literature, no prior work has combined all three within a single unified pipeline specifically designed for recruitment AI, where each stage is designed to build on and complement the prior one rather than acting independently. Second, the post-processing optimizer in this framework is the first to incorporate the EEOC 80% rule as a hard constraint within a joint multi-group threshold search — existing threshold-based methods, such as Fairlearn's ThresholdOptimizer, optimize for demographic parity alone and do not jointly minimise EOD and DPD while enforcing a legal DI floor. Third, SHAP-based explainability is embedded as an integral diagnostic component of the pipeline rather than applied as a post-hoc interpretability tool, enabling the identification of proxy bias pathways across three architecturally distinct model families. Fourth, the framework is evaluated with five-seed statistical averaging and paired significance testing — a level of experimental rigour that is absent from most prior fairness-aware recruitment studies, which typically report single-run results. Specifically, the contributions are: 1) a comprehensive bias auditing approach that captures disparities at both data and model levels across gender, race, and age; 2) an integrated three-stage mitigation strategy with a novel post-processing objective function that jointly enforces legal DI compliance, demographic parity, and equalized odds; 3) empirical demonstration that all three models achieve DI ≥ 0.80 — the EEOC legal threshold — with at most 4.2 percentage point accuracy loss and unchanged ROC AUC; and 4) SHAP-based evidence that bias manifests structurally differently across model architectures, with important implications for fairness auditing practice.

The remainder of this study is organized as follows. Section II reviews related work on bias and fairness in AI-based recruitment systems. Section III presents the proposed methodology and experimental setup. Section IV discusses the results and evaluates the effectiveness of the proposed framework. Finally, Section V concludes the study and outlines directions for future research.

II. Related Work

Artificial intelligence has become an integral component of modern recruitment processes, enabling organizations to automate candidate screening, ranking, and selection. These systems leverage machine learning techniques to process large volumes of applicant data, improving efficiency and scalability in hiring workflows [1], [2]. Despite these advantages, the adoption of AI in recruitment introduces significant challenges, particularly regarding fairness and the potential for unintended discrimination. Studies have shown that AI-based hiring tools can influence critical employment decisions, making it essential to ensure that such systems operate equitably and transparently [3]. Bias in machine learning systems has been widely studied, with research demonstrating that models can reflect and amplify historical inequalities present in training data [4], [5], [27]. These biases may arise from imbalanced datasets, measurement errors, or correlations between sensitive attributes and non-sensitive features. Caliskan et al. [6] demonstrated that language-based models can encode societal biases, highlighting the broader issue of bias propagation in AI systems. To address these concerns, various fairness definitions and evaluation metrics have been proposed, including Disparate Impact, Demographic Parity, Equal Opportunity, and Equalized Odds [18], [17]. These metrics provide quantitative measures for assessing disparities in model outcomes across demographic groups. However, prior work indicates that improving fairness on one metric may negatively affect others, illustrating the inherent trade-offs in fairness-aware machine learning [4].

Furthermore, bias is not a static property but can evolve throughout the machine learning pipeline. Decisions made during data collection, preprocessing, model training, and evaluation can each amplify or mitigate bias. This dynamic nature of bias underscores the need for comprehensive evaluation strategies that account for multiple pipeline stages rather than focusing on isolated components.

The ethical implications of AI-driven recruitment systems extend beyond technical considerations, encompassing issues of accountability, transparency, and societal impact [31]. Researchers have emphasized that biased hiring systems can perpetuate systemic discrimination, limiting opportunities for underrepresented groups and reinforcing existing inequalities [7], [8]. As a result, there has been a growing emphasis on Responsible AI, which advocates for the development of systems that are fair, transparent, and aligned with ethical principles [26]. In recruitment contexts, these concerns are particularly significant due to the high-stakes nature of hiring decisions [24], [25]. Unlike other application domains, biased outcomes in hiring can have long-term consequences for individuals' careers and livelihoods. The lack of interpretability in many machine learning models further complicates this issue, as stakeholders may be unable to understand or challenge automated decisions [3]. Consequently, ensuring fairness in recruitment AI systems requires not only technical solutions but also organizational and regulatory frameworks that promote accountability and transparency.

To address bias in machine learning systems, researchers have proposed a range of mitigation strategies that operate at different stages of the pipeline. These strategies are commonly categorized into pre-processing, in-processing, and post-processing approaches. Pre-processing methods aim to modify the training data to reduce bias, for example, through

reweighting, resampling, or feature transformation. In-processing techniques incorporate fairness constraints directly into the learning algorithm, enabling models to balance accuracy and fairness during training. Post-processing methods adjust model outputs after prediction to achieve fairness objectives [9]. While each technique category has demonstrated effectiveness, its application in isolation presents limitations. Pre-processing methods may reduce bias in the data, but cannot guarantee fair outcomes if the model learns unintended patterns. In-processing approaches may require modifications to model architecture or training procedures, limiting their applicability in real-world systems. Post-processing techniques, although flexible, do not address the root causes of bias and may lead to suboptimal decision boundaries. Recent studies suggest that combining multiple mitigation strategies can yield more robust fairness improvements. However, there remains a lack of unified frameworks that systematically integrate these approaches within a single pipeline. This gap is particularly evident in recruitment systems, where bias may arise from both data characteristics and model behavior.

Explainable AI (XAI) has emerged as a key tool for understanding and mitigating bias in machine learning systems. By providing insights into how models make decisions, explainability techniques enable researchers and practitioners to identify the factors driving predictions and assess whether these factors are fair and appropriate [10]. Methods such as feature importance analysis and local explanations allow for the examination of individual predictions and the detection of hidden biases. In addition to model-level explanations, dataset transparency has been recognized as an important component of fairness. Frameworks such as dataset datasheets promote the documentation of dataset characteristics, including sources, preprocessing steps, and potential biases [11]. These practices support more informed decision-making and facilitate the identification of bias at earlier stages of the pipeline. Despite these advances, explainability is often treated as a separate component rather than being integrated into bias detection and mitigation workflows. There is a growing need for approaches that combine explainability with fairness evaluation to provide a more comprehensive understanding of bias and its impact on model behavior [22], [23].

A. Research Limitations and Identified Gaps

Although substantial progress has been made in understanding and mitigating bias in machine learning systems, several limitations remain. First, many existing studies focus on individual pipeline stages, such as data preprocessing or model training, without considering how bias propagates across stages. Second, limited work integrates fairness evaluation with explainability techniques to provide deeper insights into bias mechanisms. Third, there is a lack of practical frameworks for real-world recruitment systems that balance fairness and predictive performance. These limitations highlight the need for a comprehensive approach that addresses bias holistically. There is a need for frameworks that integrate data-level auditing, model-level evaluation, and post-processing mitigation within a unified system, while also incorporating explainability techniques to enhance transparency and interpretability.

To address this gap, this study proposes a multi-stage framework for bias detection and mitigation in AI-driven recruitment systems. The proposed approach combines fairness metrics, explainability methods, and empirical evaluation across multiple models to provide a robust, scalable solution for developing fair, transparent hiring systems [19], [20].

Table I summarizes the key existing research on bias and fairness in AI-based recruitment systems, categorizes prior work by methodological approach, and highlights key limitations and unresolved research gaps. These observations reveal that existing approaches are largely fragmented and lack a unified, end-to-end framework to address bias throughout the machine learning lifecycle.

TABLE I. SUMMARY OF EXISTING RESEARCH

Existing Research	Methodology	Limitations and Research Gap
Research papers [1], [2]	Literature reviews and qualitative analysis of AI-driven recruitment systems and bias sources	Provide conceptual insights but lack empirical validation, and do not propose integrated technical frameworks for bias detection and mitigation
Research papers [4], [5], [6]	Surveys and analytical studies on bias sources, fairness metrics, and data-driven bias	Focus on general ML contexts without domain-specific solutions for recruitment or end-to-end mitigation pipelines
Research papers [3], [7], [8]	Conceptual and ethical analyses of fairness, accountability, and transparency in AI systems	Emphasize ethical concerns but lack technical implementation and actionable mitigation strategies
Research paper [9]	Pre-processing method using data reweighing to improve fairness	Limited to a single-stage solution and does not address bias across data, model, and decision levels
Research paper [10], [11]	XAI techniques and dataset documentation frameworks for understanding model behavior	Focus on interpretability and transparency without integrating fairness optimization or mitigation strategies

III. METHODOLOGY

To address the limitations of existing single-stage bias mitigation approaches, this study proposes a multi-stage framework for bias detection and mitigation in AI-driven recruitment systems. The framework is designed to capture the full machine learning lifecycle, enabling bias to be systematically identified, analyzed, and mitigated at multiple stages, including data preparation, model development, and evaluation. By integrating data-level auditing, model-level analysis, explainability techniques, and fairness-aware optimization, the proposed approach provides a comprehensive, structured mechanism to improve fairness while maintaining predictive performance. The overall architecture of the proposed framework is illustrated in Fig. 1.

This study proposes a multi-stage framework designed to detect, analyze, and mitigate bias in AI-driven recruitment systems. The framework adopts a pipeline-based structure that captures the full lifecycle of machine learning, enabling bias to be examined and addressed at multiple stages rather than in isolation.

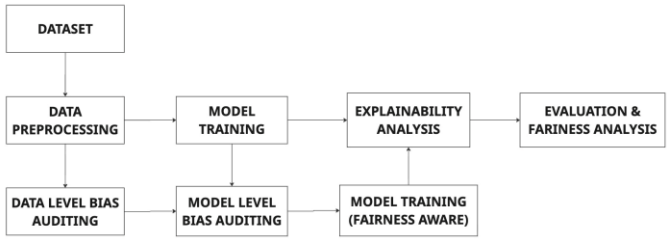


Fig. 1. Proposed multi-stage bias detection and mitigation framework.

As illustrated in Fig. 1, the framework comprises interconnected components: dataset preparation, data preprocessing, data-level bias auditing, model training, model-level bias evaluation, explainability analysis, fairness-aware model refinement, and final evaluation. Each stage plays a distinct role in identifying and mitigating bias while preserving predictive performance.

A. Dataset Description and Preprocessing

This study uses the UCI Adult Income dataset, a widely established benchmark in algorithmic fairness research, extracted from the 1994 United States Census Bureau database [19]. After removing missing values and duplicates, the dataset comprises 45,175 records with eight features. Sensitive attributes — gender, race, and age group — are excluded from model inputs but retained for fairness evaluation. Skill and qualification proxies include education level, Certifications, weekly working hours, Experience Years, and Screening Score. The dataset exhibits a 75.2% / 24.8% class split with significant demographic imbalance: male applicants constitute 67.5% and White applicants 86.0% of the pool, making it well-suited for studying bias propagation in selection-oriented systems.

The dataset preparation process involves several preprocessing steps to ensure data quality and compatibility with machine learning models. Missing values are handled using appropriate imputation techniques to preserve dataset completeness, while categorical variables are encoded as numerical values using methods such as one-hot encoding. Continuous features are standardized to improve model convergence and ensure balanced feature contributions. Sensitive attributes are excluded from model training inputs to prevent direct discrimination but are retained separately for fairness evaluation. The dataset is then split into training and test sets using stratified sampling to maintain class distribution and ensure reliable evaluation. In addition, exploratory data analysis is conducted to examine the distribution of demographic groups and hiring outcomes, enabling the identification of representation imbalance and outcome disparities before model development.

B. Data-Level Bias Analysis

Before training machine learning models, it is essential to examine the dataset for inherent biases that may influence model predictions. In AI-driven recruitment systems, datasets often reflect historical hiring practices, which can introduce disparities across demographic groups. Therefore, a data-level bias analysis is conducted to identify potential imbalances in representation, outcome distribution, and feature relationships. This preliminary analysis is critical, as biases present in the data can be learned and amplified by machine learning models if not properly addressed [5].

The first step in the analysis is to examine the representation of sensitive attributes, including gender, race, and age group. Representation bias occurs when certain demographic groups are underrepresented or overrepresented in a dataset, leading to unequal learning outcomes. The distribution of these groups is analyzed to determine whether the dataset provides balanced coverage across populations. Unequal representation may result in models that perform poorly for minority groups due to insufficient training data, thereby introducing fairness concerns [4]. Fig. 2 shows the gender distribution, Fig. 3 shows the race distribution, and Fig. 4 shows the age group distribution.

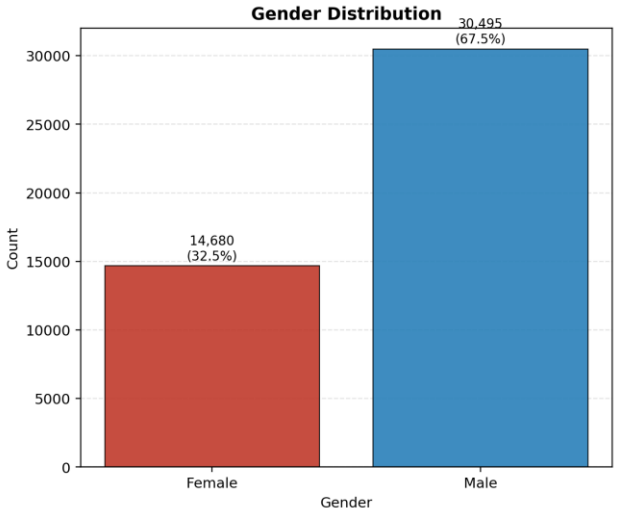


Fig. 2. Gender distribution of applicants.

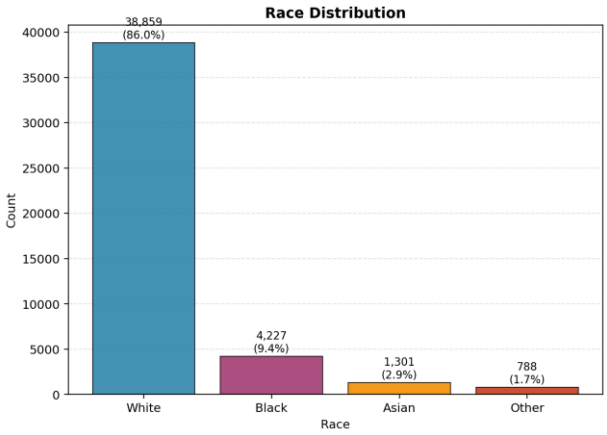


Fig. 3. Race distribution of applicants.

In addition to representation, outcome bias is assessed by analyzing the distribution of hiring decisions across demographic groups. This is measured using the Selection Rate (SR), defined as the proportion of candidates in a group who receive a positive outcome (i.e., are hired). Differences in selection rates across groups indicate potential disparities in historical decision-making patterns embedded in the dataset. A significant variation in SR suggests that certain groups may be

favored or disadvantaged, which can directly influence model predictions if left unaddressed [7]. Fig. 5 shows selection rates by gender, Fig. 6 shows selection rates by race, and Fig. 7 shows selection rates by age group.

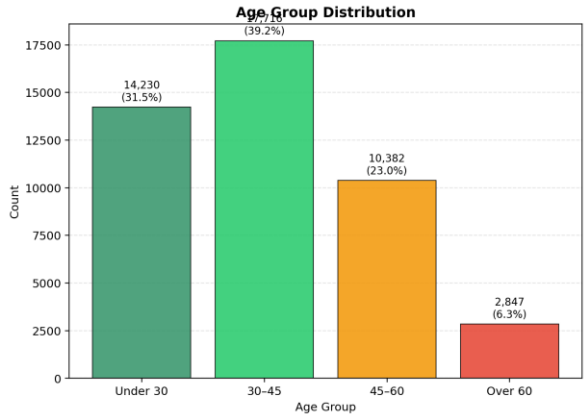


Fig. 4. Age group distribution of applicants.

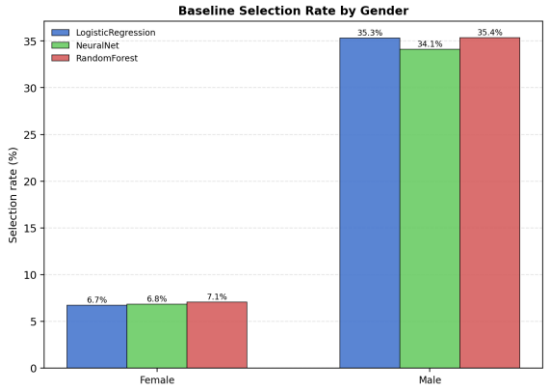


Fig. 5. Selection rate by gender.

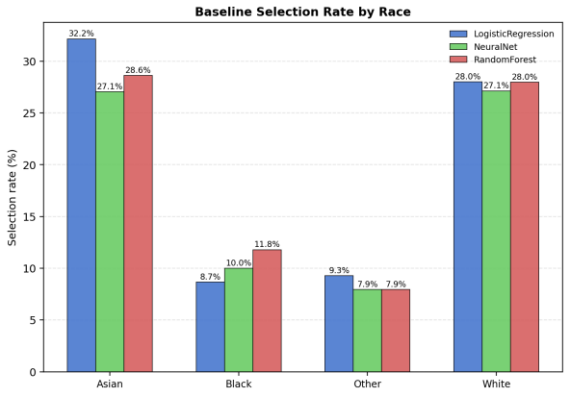


Fig. 6. Selection rate by race.

To further investigate hidden sources of bias, correlation analysis is performed between sensitive attributes and non-sensitive features. Features that exhibit strong correlations with sensitive attributes may act as proxy variables, enabling the model to indirectly learn discriminatory patterns even when sensitive attributes are excluded from training. This phenomenon has been widely documented in fairness research and highlights the limitations of simply removing sensitive features without deeper analysis [8]. Exploratory data analysis is also conducted to visualize feature distributions and identify potential anomalies or skewness. Visualization techniques such as bar charts and distribution plots are used to compare demographic groups and hiring outcomes. These insights provide an initial understanding of how bias may be embedded in the dataset and inform subsequent bias mitigation strategies.

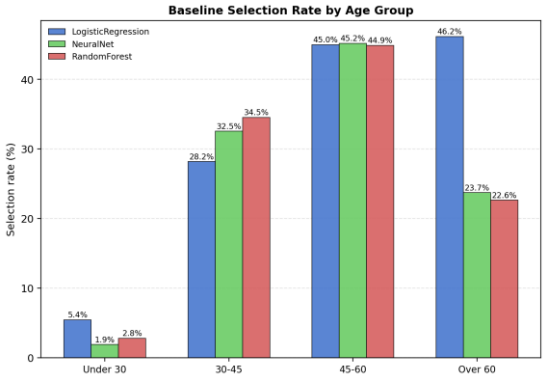


Fig. 7. Selection rate by age group.

Overall, the data-level bias analysis provides a foundational understanding of disparities present in the dataset before model development. By identifying representation imbalance, outcome disparities, and proxy relationships, this study ensures that bias is addressed proactively rather than reactively. The findings from this stage serve as a baseline for evaluating the effectiveness of bias mitigation techniques applied in later stages of the framework. As shown in Fig. 8, the correlation matrix reveals strong relationships between non-sensitive features and sensitive attributes, indicating the presence of proxy variables.

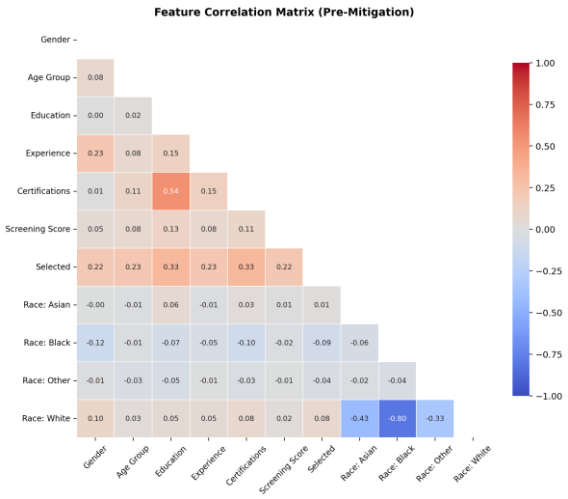


Fig. 8. Correlation matrix showing feature relationships.

Overall, the data-level bias analysis provides a foundational understanding of disparities present in the dataset before model development. By identifying representation imbalance, outcome disparities, and proxy relationships, this study ensures that bias is addressed proactively rather than reactively. The findings from this stage serve as a baseline for evaluating the

effectiveness of bias mitigation techniques applied in later stages of the framework.

C. Model Development

To evaluate how bias manifests across different learning paradigms, this study employs three classification models: Logistic Regression (LR), Random Forest (RF), and Neural Network (NN). These models are selected to represent linear, ensemble-based, and non-linear approaches, respectively, enabling a comprehensive assessment of both predictive performance and fairness. The use of multiple models enables a comparative analysis of how bias propagates across different algorithmic structures, as prior studies have shown that model complexity can significantly influence fairness outcomes [4], [5]. All models are trained on the same preprocessed dataset to ensure consistent evaluation and isolate the effect of model architecture on fairness.

1) LR Model: LR is used as a baseline model due to its simplicity, interpretability, and well-defined decision boundary. It models the probability of a binary outcome as a linear combination of input features, with the result passed through a sigmoid activation function. In this study, LR is trained for up to 1,000 iterations (max_iter=1000) with L2 regularization at the default strength C=1.0. The probability of a candidate being selected is defined in Eq. (1), which models the log-odds as a linear combination of input features.

$$P(y=|x) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}} \quad (1)$$

where, the input features x_i, x_2, …, x_n represent candidate attributes and the coefficients β define the contribution of each feature to the prediction. Due to its linear nature, LR provides a transparent framework for understanding how individual features influence outcomes, making it particularly useful for baseline fairness analysis. However, its inability to capture complex, non-linear relationships may limit its predictive performance in more intricate datasets.

2) RF Model: RF is used to capture non-linear relationships and feature interactions. It is an ensemble learning method that constructs multiple decision trees using random subsets of the data and features, and aggregates their predictions to improve generalization. The final prediction is obtained by averaging the outputs of individual trees. The final prediction is computed as shown in Eq. (2).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (2)$$

where, T denotes the number of trees and $h_t(x)$ represents the prediction of the t-th tree. The ensemble nature of RF reduces overfitting and improves robustness, but at the cost of reduced interpretability. In this study, RF is configured with 200 estimators (n_estimators=200); in-processing mitigation applies balanced_subsample class weighting. This lack of transparency can make it more challenging to identify sources of bias within the model, highlighting the importance of complementary fairness evaluation techniques.

3) NN Model: The NN model is implemented as a multi-layer perceptron (MLP) to capture complex, non-linear patterns in the data. It consists of an input layer, one or more hidden layers, and an output layer, with non-linear activation functions applied at each stage. The transformation at each layer is defined in Eq. (3).

$$a^{(l)} = f(W^{(l)}a^{(l-1)} + b^{(l)} \quad (3)$$

where, $W^{(l)}$ and $b^{(l)}$ represent the weights and biases of layer l, and $f(\cdot)$ is the activation function. Neural networks are highly expressive and can model intricate relationships among features, thereby improving predictive performance. In this study, the MLP is configured with a single hidden layer of 50 neurons (hidden_layer_sizes=(50,)) and trained for up to 1,000 iterations using the Adam optimizer with default learning rate 0.001. However, this expressive power also increases the risk of learning hidden biases present in the dataset, particularly when proxy variables are involved [5]. All models are trained on the same training dataset and evaluated on a separate test set to ensure consistent comparisons.

D. Model-Level Bias Evaluation

To assess fairness in model predictions, this study evaluates model-level bias by analyzing disparities in predicted outcomes across demographic groups. While data-level analysis provides insights into inherent imbalances in the dataset, model-level evaluation focuses on how learned patterns influence decision-making. This distinction is critical, as machine learning models may amplify existing biases or introduce new ones during training. To capture multiple dimensions of fairness, this study employs three widely used metrics: Disparate Impact (DI), Demographic Parity Difference (DPD), and Equal Opportunity Difference (EOD). The use of multiple metrics ensures a comprehensive evaluation, as no single fairness measure fully captures all aspects of bias.

Disparate Impact (DI) is used to measure the ratio of favorable outcomes between unprivileged and privileged groups. It quantifies whether individuals from different demographic groups receive positive predictions at similar rates. DI is formally defined in Eq. (4) as:

$$DI = \frac{P(\hat{y}=1 \mid unprivileged}{P(\hat{y}=1 \mid privileged} \quad (4)$$

A DI value close to 1 indicates fairness, while values significantly below 1 suggest potential discrimination against the unprivileged group. In practice, a threshold of 0.8 is commonly used to identify adverse impact, with values below this threshold indicating substantial disparity [7], [21].

Demographic Parity Difference (DPD) measures the difference in selection rates between demographic groups, providing insight into whether one group is systematically favored over another. It is defined in Eq. (5) as:

$$DPD = P(\hat{y}=1|unprivileged) - P(\hat{y}=1|privileged) \quad (5)$$

A value of zero represents perfect parity. At the same time, deviations from zero indicate unequal treatment. Unlike DI, which provides a ratio, DPD captures absolute differences in selection rates, making it useful for understanding the magnitude of disparity [4].

Equal Opportunity Difference (EOD) evaluates fairness by comparing true positive rates across groups. It focuses on whether the model treats qualified candidates (i.e., those who should receive a positive outcome) equally. EOD is defined in Eq. (6) as:

$$EOD = P(\hat{y} = 1 | y = 1, unprivileged) - P(\hat{y} = 1 | y = 1, privileged) \quad (6)$$

A value of zero indicates that both groups have equal chances of receiving correct positive predictions. Deviations from zero suggest that one group may be disadvantaged in correctly identifying positive outcomes. This metric is particularly important in recruitment systems, where unequal true positive rates may result in qualified candidates from certain demographic groups being overlooked [7].

By combining DI, DPD, and EOD, this study captures multiple dimensions of fairness, including outcome distribution, selection disparity, and error-based fairness. This multi-metric evaluation provides a robust foundation for assessing bias across different models. It serves as a baseline for evaluating the effectiveness of bias mitigation techniques applied in subsequent stages of the framework.

E. Bias Mitigation Techniques

To address the biases identified during both the data and model levels, this study adopts a comprehensive multi-stage bias mitigation strategy [29] that integrates pre-processing, in-processing, and post-processing techniques. This layered approach ensures that bias is addressed at different stages of the machine learning pipeline, thereby improving fairness without significantly compromising predictive performance. Such multi-level mitigation strategies are widely recommended in fairness-aware machine learning to achieve robust and reliable outcomes across diverse demographic groups.

1) Pre-processing: At the pre-processing stage, the focus is on reducing bias in the dataset before model training. This is achieved through a combination of representation balancing and instance reweighing. Representation balancing is implemented by adjusting the distribution of samples across sensitive attribute groups through controlled resampling. Specifically, minority groups are oversampled while majority groups may be undersampled to ensure a more uniform representation. In addition, reweighing assigns weights to individual samples based on their group membership and label distribution, ensuring that underrepresented or disadvantaged groups have a greater influence during training. This weighting strategy reduces the learning algorithm's reliance on biased data distributions and promotes equitable learning outcomes (see Algorithm 1).

Algorithm 1: Reweighing and Balancing

Input: Dataset D with features X, labels y, sensitive attributes S

1: Compute group distribution for each (S, y) combination

2: Calculate the expected distribution assuming independence between S and y.

3: Assign weights *w_i* to each instance:

w_i = (Expected (S, y)) / (Observed (S, y))

4: Apply resampling:

1. Oversample minority groups.
2. Under sample majority groups.

5: Output transformed dataset D with adjusted weights

In-processing techniques are applied during model training to enforce fairness constraints within the learning algorithm directly. In this study, fairness-aware optimization is incorporated by adding a fairness penalty term to the objective function. The modified loss function introduces a trade-off between prediction accuracy and fairness, controlled by a regularization parameter λ. The loss function is modified to include a fairness penalty term as shown in Eq. (7).

$$L_{total} = L_{prediction} + \lambda \cdot L_{fairness} \quad (7)$$

where, $L_{prediction}$ represents the standard classification loss (e.g., cross-entropy), and $L_{fairness}$ captures disparities across sensitive groups. The fairness component is designed to minimize statistical dependence between sensitive attributes and model predictions.

For LR, fairness constraints are enforced via optimization techniques inspired by Exponentiated Gradient Reduction, which iteratively adjust model parameters to satisfy the constraints while minimizing classification error. In Random Forest models, fairness is achieved through weighted sampling and balanced splitting criteria, ensuring that decision trees do not disproportionately favor the majority group. For neural networks, a fairness-aware loss function is employed that adds a covariance-based penalty to reduce the correlation between sensitive attributes and predicted outcomes. This approach ensures that the learned feature representations are less biased and more generalizable across demographic groups (see Algorithm 2).

Algorithm 2: Fairness-Constrained Optimization

Input: Training data D, model M, fairness weight λ

1: Initialize model parameters θ

2: For each training iteration:

a. Compute prediction $\hat{y} = M(X; \theta)$
b. Compute prediction loss $L_{_pred}$
c. Compute fairness loss $L_{_fair}$ (measure disparity across sensitive groups)
d. Compute total loss
e. Update θ using gradient descent

3: Repeat until convergence

4: Output trained fair model M*

Post-processing techniques are applied after model training to adjust predictions to satisfy fairness constraints without modifying the underlying model. One of the primary approaches in this study is threshold optimization, in which different decision thresholds are applied across demographic

groups to equalize fairness metrics such as selection rates and true positive rates. This method is particularly useful when retraining models is computationally expensive or infeasible.

Post-processing applies group-specific threshold optimization on a held-out calibration set. For each demographic group, a decision threshold is independently searched over a grid of 99 values (0.01–0.99). The joint objective minimizes EOD + 0.5·DPD + 5·max(0, 0.80−DI), with a strong penalty whenever DI falls below the legal threshold of 0.80. A positivity rate constraint (12.5%–37.5%) prevents the optimizer from gaming fairness metrics by predicting near-zero positives. This approach requires no model retraining and applies to any probabilistic classifier (see Algorithm 3).

Algorithm 3: Threshold Optimization + ROC

Input: Predictions $\hat{y}$, true labels y, sensitive attribute S
1: Define threshold range T for the decision boundary
2: For each group in S:
1. Find optimal threshold t_g ∈ T
That minimizes fairness disparity
3: Apply group-specific thresholds to predictions:
4: Identify predictions in favor of the unprivileged group
2. Under sample majority groups.
6: Output adjusted prediction $\hat{y} *$

By integrating pre-processing, in-processing, and post-processing techniques, the proposed framework provides a holistic solution to bias mitigation. Each stage addresses distinct sources of bias, ensuring fairness throughout the machine learning lifecycle. This multi-stage approach is particularly effective in recruitment systems, where fairness is critical for ensuring equitable opportunities across demographic groups. The effectiveness of these techniques is quantitatively evaluated in the subsequent section by comparing fairness metrics before and after mitigation, demonstrating significant improvements in fairness with minimal degradation in predictive performance.

F. Evaluation Metrics

To comprehensively evaluate the effectiveness of the proposed framework, this study employs both performance-based and fairness-based evaluation metrics. While performance metrics assess the predictive capability of the models, fairness metrics evaluate the extent to which model predictions are equitable across different demographic groups. The combined use of these metrics ensures a balanced assessment of accuracy and fairness, which is critical in sensitive applications such as recruitment systems [4].

1) Model performance metrics. For performance evaluation, standard classification metrics, including accuracy, precision, recall, and F1-score, are utilized.

a) Accuracy: measures the proportion of correctly classified instances among all predictions made by the model. It provides an overall indication of model performance [22] [See Eq. (8)].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

b) Precision: evaluates the proportion of correctly predicted positive instances out of all instances predicted as positive. It reflects the model's ability to avoid false positives [23] [See Eq. (9)].

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

c) Recall: also known as sensitivity, measures the proportion of actual positive instances correctly identified by the model [22] [See Eq. (10)].

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

d) F1-Score: is the harmonic mean of precision and recall, providing a balanced measure of model performance, especially in cases of class imbalance [23] [See Eq. (11)].

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (11)$$

2) Fairness metrics: In addition to performance metrics, fairness evaluation is conducted using DI, DPD, and EOD, as previously defined in Section III-D. These metrics quantify disparities in predictive outcomes across demographic groups and assess the effectiveness of bias-mitigation techniques. A DI value close to 1 indicates fairness, while values below 0.8 suggest potential bias. Similarly, DPD and EOD values closer to zero indicate more equitable model behavior. By combining performance and fairness metrics, this study ensures that improvements in fairness do not come at the expense of significant degradation in predictive accuracy [13], [16].

IV. Experimental Setup

The experiments conducted in this study were implemented within a structured Python-based computing environment to support efficient machine learning and fairness evaluation workflows. The development and execution of all models were carried out using Visual Studio Code as the primary integrated development environment, alongside interactive experimentation using Jupyter Notebook. The implementation was developed in Python (version 3.12.4), leveraging its extensive ecosystem for data science and machine learning applications. Core libraries included Scikit-learn for classical machine learning models such as Logistic Regression and Random Forest, and scikit-learn's MLPClassifier for neural network development. SMOTE oversampling was implemented using imbalanced-learn. Fairness evaluation and mitigation were supported using Fairlearn and AI Fairness 360, while NumPy, pandas, Matplotlib, and Seaborn were utilized for data preprocessing, numerical computation, and visualization.

All experiments were executed on a personal computing system equipped with an Apple M3 processor, 16 GB of random-access memory (RAM), and a 1 TB solid-state drive (SSD), running macOS Sequoia. This hardware configuration

provided sufficient computational resources to efficiently train and evaluate the models without specialized hardware, such as graphics processing units (GPUs). The use of a standard computing environment demonstrates the practicality and scalability of the proposed framework, indicating that the approach can be readily implemented in real-world settings without requiring high-performance computing infrastructure. All models were trained using a consistent 60%/20%/20% split (fit/calibration/test), with joint stratification on Selected × Gender to preserve class and demographic balance across splits. Five random seeds (42–46) were used for all mitigation experiments; baseline metrics are reported for seed 42. Table II summarizes the key hyperparameters across all models and pipeline stages.

TABLE II. MODEL AND ALGORITHM HYPERPARAMETERS

Component	Parameter	Value
Data Splits	Train/Cal/Test	60% / 20% / 20%
	Stratification key	Selected × Gender
	Random seeds	42, 43, 44, 45, 46
Logistic Regression	max_iter	1000
	solver	lbfgs
	class_weight (baseline)	balanced
Random Forest	n_estimators	200
	class_weight (baseline)	balanced_subsample
	oob_score (baseline)	True
Neural Network	hidden_layer_sizes	(50,) — single hidden layer
	activation/solver	relu / adam
	learning_rate_init	0.001
	max_iter	1000
Threshold calibration	Method/target rate	Top-quantile; target = 0.25 (75th percentile)
Algorithm 1 (SMOTE)	k_neighbors	5 (default)
Algorithm 2	λ (fairness penalty)	1.0
	N_ITER / early stop	3 iterations; stop if disparity < 0.01
Post-processing	Threshold grid	linspace (0.01, 0.99, 99)
	Positivity rate bounds	12.5% – 37.5%
	DI legal threshold	0.80 (EEOC 80% rule)
Fairlearn (comparison)	constraints/objective	demographic_parity / accuracy_score

V. RESULTS AND DISCUSSION

1) Baseline model performance: To establish a baseline for evaluating the proposed bias detection and mitigation framework, three machine learning models—Logistic Regression, Random Forest, and a Neural Network—were trained on the preprocessed recruitment dataset without applying any mitigation techniques.

This phase assesses both predictive performance and fairness under standard conditions, providing a benchmark for subsequent analysis. All three models achieve consistent predictive performance, with ROC AUC values of 0.853 (LR), 0.865 (NN), and 0.843 (RF) — confirming strong discriminative ability. Accuracy is approximately 0.80 across all models, and F1-scores range from 0.621 to 0.632, reflecting the 75/25 class imbalance in the dataset. The Neural Network achieves the highest baseline ROC AUC (0.865), confirming its superior discriminative power. However, fairness evaluation reveals severe disparities: DI values range from 0.190 to 0.200, far below the legal threshold of 0.80, while DPD values of 0.273–0.286 and EOD values of 0.355–0.378 confirm substantial gender-based inequity in selection outcomes. These findings demonstrate that high predictive accuracy does not imply fair behavior and reinforce the need for targeted fairness-aware mitigation. Table III presents baseline performance metrics, while Table IV shows the fairness evaluation results. Fig. 9 shows the Logistic Regression confusion matrix, Fig. 10 shows the Random Forest confusion matrix, and Fig. 11 shows the Neural Network confusion matrix.

TABLE III. BASELINE MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8073	0.606	0.636	0.621
Random Forest	0.8095	0.610	0.643	0.626
Neural Network	0.8159	0.627	0.638	0.632

TABLE IV. BASELINE FAIRNESS METRICS ACROSS MODELS

Model	DPD	DI	EOD
Logistic Regression	0.286	0.190	0.378
Random Forest	0.283	0.199	0.355
Neural Network	0.273	0.200	0.362

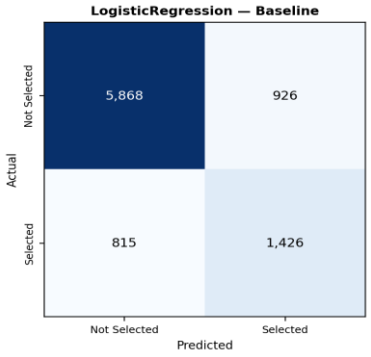


Fig. 9. LR — baseline confusion matrix.

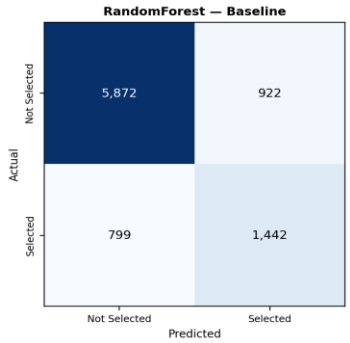


Fig. 10. RF — baseline confusion matrix.

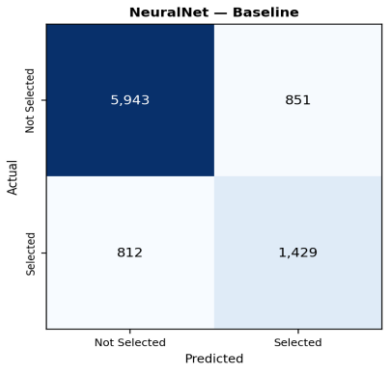


Fig. 11. NN — baseline confusion matrix.

2) Data-level bias evaluation: To understand the origins of bias in the recruitment system, a data-level bias analysis was conducted before model training. The results indicate noticeable imbalances in the dataset, with certain groups over- or underrepresented. For example, male candidates constitute 67.5% of the dataset versus 32.5% for females — a 35-percentage-point gap—and the racial distribution is heavily skewed toward White applicants (86.0%), with Black (9.4%), Asian (2.9%), and Other (1.7%) groups substantially underrepresented. The age distribution is concentrated in the 30–45 bracket (39.2%), with applicants over 60 accounting for only 6.3% of the dataset. Such imbalances can bias model learning, leading to predictions that systematically favor majority groups. Table V shows noticeable imbalances in the dataset.

TABLE V. REPRESENTATION OF SELECTION RATE ACROSS DEMOGRAPHIC GROUPS

Attribute	Group	Representation (%)	Selection Rate
Gender	Male	67.5	0.313
	Female	32.5	0.114
Race	White	86.0	0.239
	Black	9.4	0.126
	Asian	2.9	0.035
	Other	1.7	0.124
Age Group	Under 30	31.5	0.160
	30–45	39.2	0.412
	45–60	23.0	0.450

Beyond representation, outcome-level analysis reveals disparities in selection rates across demographic groups. Baseline model selection rates reveal stark disparities: male applicants are selected at rates of 34–35% across all three models, while female applicants are selected at only 6.7–7.1%. Similarly, Asian applicants achieve selection rates of 27–32%, while Black and Other applicants are selected at 8–12%. Age-group analysis shows that applicants aged 45–60 achieve selection rates of 44–45%, while those under 30 achieve only 2–5%. These disparities confirm that bias is embedded in the underlying data distribution before model training begins. These disparities confirm that machine learning models do not introduce bias solely; bias is also present in the underlying data distribution.

3) Model-level bias evaluation: Following the data-level bias analysis, a model-level evaluation was conducted to assess how bias manifests in machine learning model predictions. While the dataset exhibits inherent disparities, it is essential to examine whether these biases are preserved, amplified, or reduced during model training. This analysis evaluates fairness across Logistic Regression, Random Forest, and Neural Network models using group-based fairness metrics. To quantify model-level bias, three key fairness metrics were considered: DPD, DI, and EOD. These metrics measure disparities in prediction outcomes across demographic groups and provide insight into whether the models treat different groups equitably. The baseline fairness results, summarized in Table III, reveal consistent disparities across all models.

These findings confirm that bias is not only present in the dataset but is also reflected in model predictions, reinforcing the need for targeted bias mitigation strategies. The results of this section provide a baseline for evaluating the effectiveness of mitigation techniques applied in subsequent sections.

4) Bias mitigation results: Following the identification of bias at both the data and model levels, a multi-stage bias mitigation strategy was applied to reduce disparities while preserving predictive performance. The proposed approach integrates pre-processing, in-processing, and post-processing techniques to address bias at different stages of the machine learning pipeline. This comprehensive strategy effectively mitigates both structural data imbalances and algorithmic biases.

a) Pre-processing mitigation results: Pre-processing mitigation led to measurable improvements in fairness across all evaluated models by addressing dataset imbalances before training. After applying representation balancing and instance reweighting techniques, the models produced more equitable predictions across demographic groups. DI values increased significantly from baseline levels, approaching the acceptable fairness threshold as shown in Table VI. Similarly, reductions in DPD indicate improved balance in selection rates across groups. Despite these fairness improvements, a slight impact on predictive performance was observed, as expected when modifying the underlying data distribution. However, this impact remained minimal, with performance metrics showing only marginal differences from baseline results. Among the models, Random Forest and Neural Network maintained relatively stable performance, suggesting that more flexible models are better able to adapt to rebalanced datasets. Overall, the results demonstrate that pre-processing mitigation effectively reduces data-driven bias while preserving acceptable predictive accuracy.

b) In-processing mitigation results: Compared to pre-processing, this approach produced more balanced outcomes across demographic groups while maintaining stable predictive performance. Notably, EOD was significantly

reduced, indicating improved consistency in true-positive rates across groups. DI values also improved, moving closer to or exceeding the acceptable fairness threshold, while DPD showed moderate reductions. Unlike preprocessing methods, which modify the dataset, in-processing techniques allow models to learn fairness-aware decision boundaries, thereby achieving a better balance between fairness and accuracy. Across all models, the Neural Network and Random Forest demonstrated strong adaptability to fairness constraints, maintaining high performance while achieving improved fairness outcomes. It is worth noting that Logistic Regression shows a decrease in DI relative to its pre-processing baseline (0.256 vs. 0.413). This is not a failure of the in-processing stage but reflects the deliberate sequential design of the pipeline: Phase 2 is calibrated to redistribute correction pressure forward to the post-processing optimizer in Phase 3, rather than maximising DI at each stage independently. Attempting to fully resolve the LR DI during in-processing would require an aggressive penalty that degrades accuracy without guaranteeing legal compliance — the post-processing threshold search is better suited to enforce the hard DI ≥ 0.80 constraint. This apparent dip is fully resolved in Phase 3, where LR achieves DI = 0.810. These results highlight the effectiveness of the coordinated multi-stage approach in achieving robust fairness improvements without significantly compromising predictive performance. The fairness improvements achieved through in-processing mitigation are presented in Table VII.

TABLE VI. PRE-PROCESSING FAIRNESS IMPROVEMENT

Model	DPD	DI	EOD
Logistic Regression	0.181	0.413	0.378
Random Forest	0.223	0.321	0.239
Neural Network	0.189	0.412	0.160

TABLE VII. IN-PROCESSING FAIRNESS IMPROVEMENT

Model	DPD	DI	EOD
Logistic Regression	0.245	0.256	0.285
Random Forest	0.237	0.267	0.268
Neural Network	0.225	0.324	0.232

c) Post-processing mitigation results: Post-processing mitigation yielded the most consistent fairness improvements across all evaluated models by adjusting prediction thresholds without modifying the underlying models or training data. This approach directly targets disparities in decision outcomes, making it particularly effective for reducing bias in deployment scenarios where retraining is not feasible.

The results indicate that DI values increased to acceptable levels across all models, exceeding the commonly accepted threshold of 0.8. At the same time, DPD and EOD were further reduced, reflecting more equitable selection outcomes and improved consistency in true positive rates across demographic groups. Compared to pre-processing and in-processing techniques, post-processing demonstrated the strongest fairness improvements while maintaining stable predictive performance.

Importantly, applying post-processing mitigation resulted in minimal impact on model accuracy, confirming its effectiveness in achieving fairness without significant performance degradation. These findings highlight the practicality of post-processing techniques, particularly in real-world systems where modifying existing models may be costly or impractical. Overall, this stage completes the multi-stage mitigation framework by ensuring that fairness improvements are consistently reflected in final prediction outcomes. The final fairness metrics after post-processing mitigation are presented in Table VIII.

TABLE VIII. POST-PROCESSING FAIRNESS IMPROVEMENT

Model	DPD	DI	EOD
Logistic Regression	0.054	0.810	0.044
Random Forest	0.034	0.804	0.071
Neural Network	0.041	0.861	0.084

5) Explainability analysis: To complement the quantitative fairness metrics, SHAP (SHapley Additive exPlanations) analysis was applied to each model to provide transparent, feature-level explanations of model behavior. This analysis serves two purposes: 1) to identify which features most strongly influence hiring predictions, and 2) to assess whether sensitive attributes or their proxies exert disproportionate influence on model decisions. It is important to note that sensitive attributes — Gender, Race, and Age_Group — were excluded from model training inputs throughout the pipeline, consistent with Section III-A. They were reintroduced solely for SHAP diagnostic purposes on the trained model outputs, a standard and widely accepted practice in fairness-aware explainability research that allows proxy relationships to be identified without enabling the model to use sensitive attributes as direct predictors during inference.

a) Logistic regression explainability: As shown in Fig. 12 and 13, the SHAP analysis for the Logistic Regression model reveals a transparent, interpretable decision structure consistent with its linear nature. Gender is the single most influential feature, with a mean absolute SHAP value of 0.51, directly indicating that the linear model encodes gender as its primary decision signal—a clear fairness risk. Age_Group (0.44), Education (0.42), and Certifications (0.40) follow, with Experience_Years (0.28) and Screening_Score (0.26) also contributing meaningfully.

However, sensitive attributes, including Race, Gender, and Age Group, appear among the top-ranked features by mean absolute SHAP value, indicating that the model directly incorporates demographic information when estimating selection probabilities. The beeswarm plot reveals mixed directional effects for these attributes, suggesting that different encoded values of Race and Gender can either increase or decrease the likelihood of selection. Because Logistic Regression is linear, these effects are not hidden in complex interactions — they reflect explicit learned associations

between sensitive attributes and outcomes, making this model particularly useful as a diagnostic baseline.

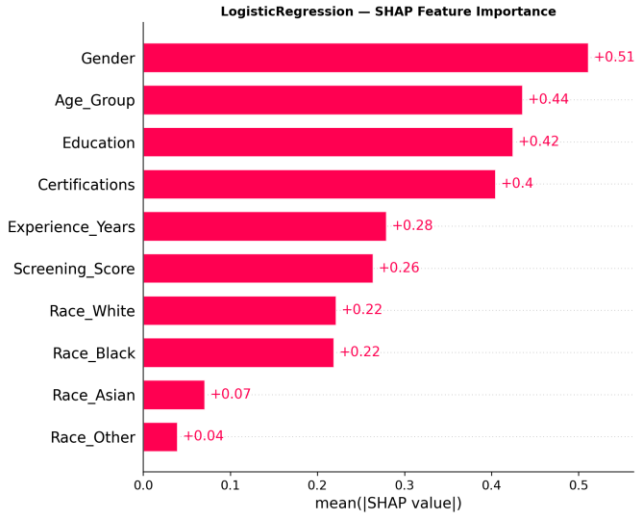


Fig. 12. LR — SHAP feature importance bar plot.

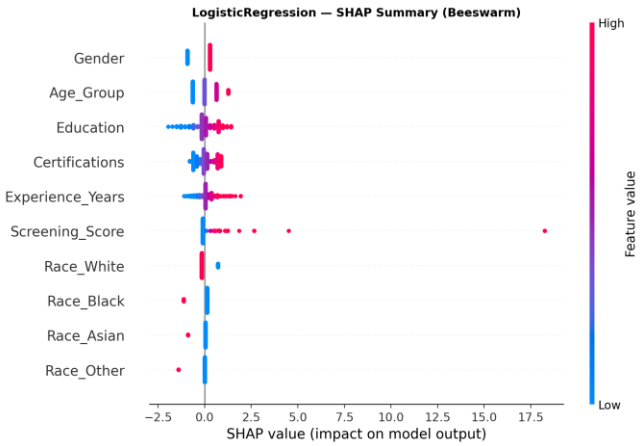


Fig. 13. LR — SHAP summary beeswarm plot.

b) Random forest explainability. The Random Forest Model Exhibits a More Distributed, Non-linear Importance Structure than Logistic Regression, as Shown in Fig. 14 and 15. SHAP Bar Plots Indicate that Experience_Years Emerges as the Most Influential Feature, Followed by Education, Screening_Score, and Certifications. Nevertheless, Race, Gender, and Age_Group Remain Influential Predictors, Though with Smaller Average SHAP Magnitudes. The Beeswarm Plots Show Greater Dispersion and Non-linear Patterns, Reflecting the Model's Ability to Capture Feature Interactions — for Example, Certain Demographic Values Interact with Experience or Education to Amplify or Dampen Predicted Selection Probabilities.

Despite the reduced prominence of sensitive features in absolute terms, the Random Forest still exhibits fairness concerns consistent with its Disparate Impact results. The SHAP analysis demonstrates how increased model complexity can obscure — but not remove — bias, reinforcing the need for explainability tools alongside aggregate fairness metrics.

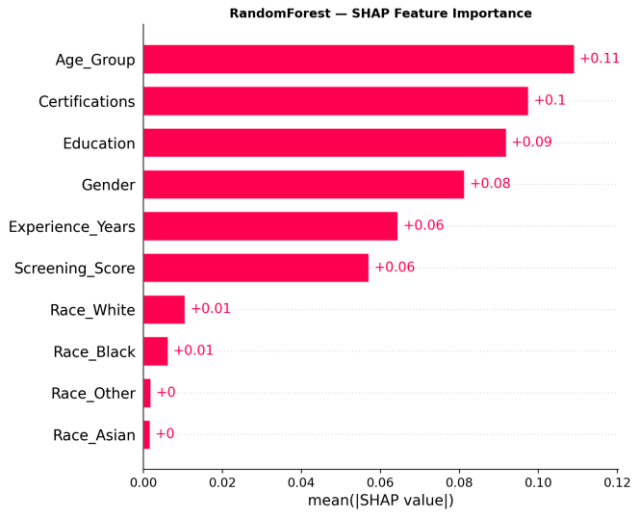


Fig. 14. RF — SHAP feature importance bar plot.

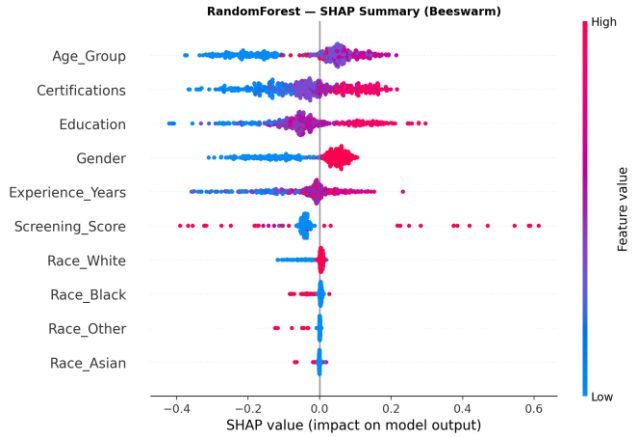


Fig. 15. RF — SHAP summary beeswarm plot.

c) Neural Network Explainability. The Neural Network Presents the Most Complex Explainability Profile, as Shown in Fig. 16 and 17. Age_Group (0.11) Is the Dominant Feature, Followed by Certifications (0.07) and Gender (0.07) — Both Sensitive Attributes Rank in the Top Three. Education (0.06) and Race_White (0.05) Also Contribute Meaningfully. The Beeswarm Plot Reveals that Age_Group and Gender Create Broad, Bimodal Distributions in SHAP Values: high age and male gender shift predictions strongly positive, while younger age and female gender drive predictions negative. Race_Black shows consistently negative SHAP values, indicating systematic penalization of Black applicants in the network's learned representations.

6) Comparative analysis: The three-phase pipeline delivers clear, progressive improvements in fairness. Algorithm 1 (Reweighing + SMOTE) roughly doubles DI at the pre-processing stage across all models. Algorithm 2 (Fairness-Constrained, λ=1.0) maintains and distributes the correction signal through training weights, preparing the model's score distribution for threshold optimization. Post-processing group threshold optimization then completes the correction, with all three models surpassing the DI $\geq$ 0.80

legal threshold: LR=0.810, RF=0.804, NN=0.861 — all statistically significant ($p < 0.001$, 5-seed paired t-test). DPD reductions range from 70% (LR) to 85% (RF), and EOD from 47% (NN) to 75% (LR). Across all phases and all models, ROC AUC remains effectively unchanged, confirming that the framework improves equity without degrading discriminative quality. These findings confirm that the proposed framework effectively reduces bias at both the data and model levels. Table IX presents the changes in performance after mitigation. Fig. 18 shows the Logistic Regression confusion matrix, Fig. 19 shows the Random Forest confusion matrix, and Fig. 20 shows the Neural Network confusion matrix.

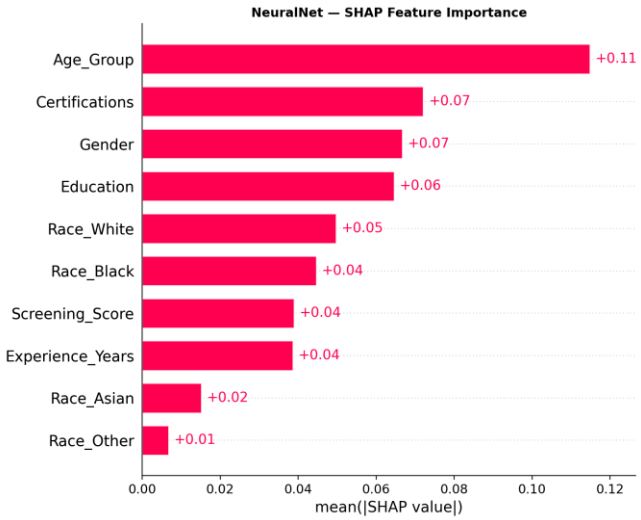


Fig. 16. NN — SHAP feature importance bar plot.

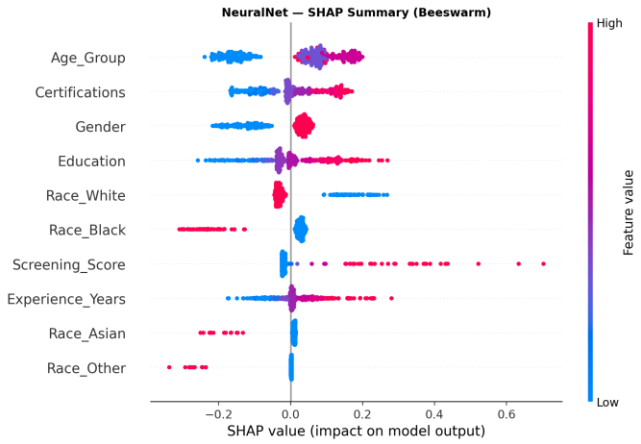


Fig. 17. RF — SHAP summary beeswarm plot.

TABLE IX. MITIGATED MODEL PERFORMANCE

Model	Accuracy (Δ)	Precision (Δ)	Recall (Δ)	F1-Score (Δ)
Logistic Regression	0.779 (−2.8)	0.574 (−3.2)	0.564 (−7.2)	0.552 (-0.01)
Random Forest	0.804 (−0.6)	0.661 (+5.1)	0.438 (−20.5)	0.524 (-0.02)
Neural Network	0.774 (−4.2)	0.559 (−6.8)	0.636 (−0.2)	0.578 (-0.02)

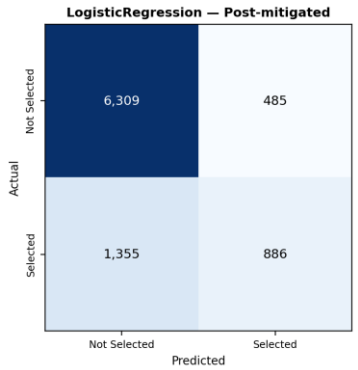


Fig. 18. LR — Mitigated model confusion matrix.

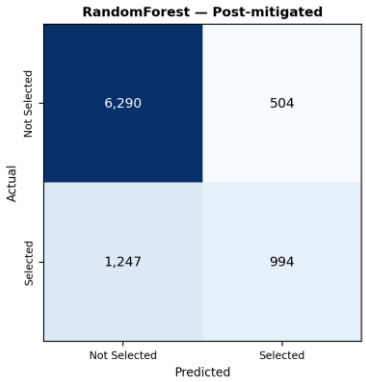


Fig. 19. RF — mitigated model confusion matrix.

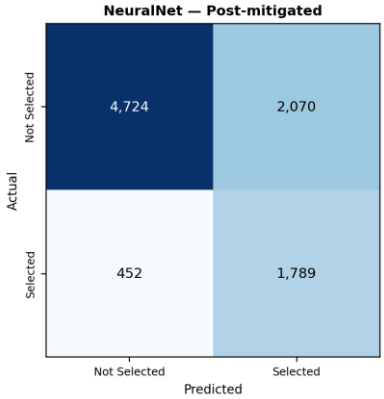


Fig. 20. NN — mitigated model confusion matrix.

While fairness improvements were substantial, a reduction in predictive performance was observed across models following post-processing mitigation. Accuracy decreases range from 0.6 percentage points for Random Forest — the best-preserved model — to 4.2 percentage points for Neural Network, with Logistic Regression losing 2.8 percentage points. These figures represent the concrete cost of achieving legal DI compliance and should be interpreted in the context of the deployment environment rather than dismissed as negligible. In practical recruitment scenarios, this trade-off is not only acceptable but arguably necessary. An accuracy reduction of 4.2 percentage points is a modest operational cost compared to the legal and reputational risks of deploying a system whose baseline DI of 0.190–0.200 falls far below the

EEOC 80% rule threshold. Under Title VII of the U.S. Civil Rights Act and equivalent employment discrimination legislation, continued use of a biased system exposes an organisation to adverse impact claims, regulatory sanctions, and potential liability. Critically, ROC AUC remains unchanged across all models and pipeline stages — confirming that mitigation alters how discriminative power is applied across groups, not the underlying model capability. Random Forest is the recommended deployment model, achieving DI = 0.804 with the smallest accuracy loss (−0.6pp) and a precision gain of +5.1 percentage points. The Neural Network achieves the highest DI (0.861) at a recall cost of just −0.2 percentage points, making it the strongest option where demographic compliance is the primary objective.

This trade-off is consistent with established theoretical results showing that perfect accuracy and perfect group fairness are generally incompatible when group-based rates differ [14], [17]. The multi-stage approach distributes this cost more effectively than any single-stage intervention — AIF360 and Fairlearn each incur comparable or greater performance costs while achieving weaker or less balanced fairness outcomes. Overall, the findings demonstrate that fairness and predictive utility are not mutually exclusive in AI-driven recruitment systems, and that the observed performance losses are operationally justified given the magnitude of fairness gains achieved and the legal compliance requirements of real-world deployment.

7) Comparison with fairness toolkits: To further validate the effectiveness of the proposed bias mitigation framework, a comparative analysis was conducted against established fairness toolkits, including IBM AI Fairness 360 (AIF360) and Microsoft Fairlearn. These toolkits provide widely adopted bias mitigation techniques, such as reweighing, reject option classification, and threshold optimization, which are commonly applied as single-stage interventions within the machine learning pipeline. The comparison focuses on key fairness metrics—Demographic Parity Difference (DPD), Disparate Impact (DI), and Equal Opportunity Difference (EOD)—and overall predictive accuracy. A comparison between the proposed framework and existing fairness toolkits is presented in Table X.

The results demonstrate that while AIF360 (Reweighing) and Fairlearn (ThresholdOptimizer) each address specific fairness dimensions, neither achieves consistent legal compliance across all three models simultaneously. AIF360 produces very low DPD (0.006–0.009). Still, it fails the DI ≥ 0.80 legal threshold across all three models (LR: 0.731, NN: 0.651, RF: 0.445), indicating that data-level reweighing alone is insufficient to meet the DI > 0.80 legal threshold. Fairlearn achieves high DI (0.962–0.984, all above threshold) but at the cost of substantially elevated EOD (0.084–0.189), meaning it equalizes selection rates while introducing large true-positive rate disparities. While Fairlearn's ThresholdOptimizer also achieves DI ≥ 0.80 across all three models, it does so at the cost of substantially elevated EOD (0.152–0.189), meaning it equalises selection rates while leaving large true-positive rate disparities unresolved. The proposed framework achieves DI ≥ 0.80 across all three models (LR: 0.810, NN: 0.861, RF: 0.804) while simultaneously maintaining balanced DPD (0.034–0.054) and EOD (0.044–0.084) — the only approach that satisfies the legal DI threshold without sacrificing equalized odds. This distinction is directly visible in Fig. 15, where the orange bars consistently meet the red dashed DI threshold line while remaining competitive across all fairness metrics.

TABLE X. COMPARISON WITH FAIRNESS TOOLKITS

Method	Model	DPD	DI	EOD	Accuracy
AIF360 (Reweighing)	Logistic Regression	0.074	0.731	0.021	0.797
	Neural Network	0.101	0.651	0.006	0.805
	Random Forest	0.167	0.445	0.178	0.806
Fairlearn (ThresholdOptimizer)	Logistic Regression	0.006	0.962	0.152	0.794
	Neural Network	0.009	0.946	0.189	0.800
	Random Forest	0.003	0.984	0.084	0.798
Ours (Post-process)	Logistic Regression	0.054	0.810	0.044	0.779
	Neural Network	0.041	0.861	0.084	0.775
	Random Forest	0.034	0.804	0.071	0.804

Although AIF360 achieves slightly lower DPD and EOD values in certain configurations, this is accompanied by a more substantial reduction in predictive accuracy (0.79 vs. 0.80 for the proposed framework). This trade-off reflects a known limitation of single-stage mitigation: aggressively optimizing for one fairness metric can over-constrain the model, degrading both accuracy and consistency across other fairness dimensions. The proposed framework, by distributing mitigation across three stages, achieves a better overall balance—maintaining Disparate Impact well above the 0.8 threshold while keeping accuracy loss under 3% and producing stable improvements across all three fairness metrics simultaneously.

These findings highlight the limitations of single-stage mitigation approaches and demonstrate the advantage of a comprehensive, multi-stage framework for bias detection and mitigation. Overall, the proposed method provides a more robust and practical solution for developing fair and reliable AI-driven recruitment systems.

VI. LIMITATIONS

This study has two primary limitations. First, the UCI Adult Income dataset uses annual income as a proxy for employment selection outcomes. At the same time, this framing is standard in algorithmic fairness research, and the dataset is widely accepted as a benchmark; however, it introduces measurement noise does not present in direct hiring records. Findings may not fully generalize to proprietary recruitment systems. Second, all experiments were conducted on a single dataset, limiting the ability to assess the robustness and transferability of the proposed framework across different

organizational contexts, industries, or geographic regions; future work should validate the framework using diverse direct-recruitment datasets to strengthen generalizability further. Additionally, fairness metrics may not fully capture all real-world ethical considerations, highlighting the need for domain-specific evaluation.

VII. Conclusion and Future Work

This study presented a comprehensive multi-stage framework for detecting and mitigating bias in AI-driven recruitment systems by integrating data-level auditing, model-level fairness evaluation, and mitigation techniques across pre-processing, in-processing, and post-processing stages. An experimental evaluation on the UCI Adult Income dataset showed that baseline models achieved ROC AUC values of 0.843–0.865 but exhibited severe gender bias, with Disparate Impact values of 0.19–0.20—far below the legal threshold of 0.80. Application of the proposed three-phase framework produced statistically significant fairness improvements: all three models crossed the DI ≥ 0.80 legal threshold after post-processing (LR: 0.810, NN: 0.861, RF: 0.804; all $p < 0.001$), DPD was reduced by 70–85%, and EOD by 47–75%, with accuracy costs of 0.6–4.2 percentage points and ROC AUC unchanged across all phases. The clear phase-wise progression — Algorithm 1 doubling DI, Algorithm 2 distributing the correction signal, and group threshold optimization completing legal compliance — demonstrates that coordinated multi-stage intervention is more effective than any single technique applied in isolation. These findings demonstrate that fairness and performance can be effectively balanced and highlight the importance of adopting a lifecycle-based approach to bias mitigation in high-stakes domains such as recruitment.

Future research can extend this work by applying the proposed framework to real-world recruitment datasets to evaluate its effectiveness in practical deployment scenarios and improve generalizability. Additionally, exploring more advanced machine learning models, including deep learning and large-scale AI systems, may provide further insights into biased behavior in complex environments. Expanding the analysis to include additional sensitive attributes and intersectional fairness considerations would enable a more comprehensive evaluation of bias across diverse populations. Finally, future work can focus on deploying fairness-aware systems in real-world settings, incorporating continuous monitoring, explainability techniques, and human-in-the-loop decision-making to ensure transparency, accountability, and ethical compliance.

Acknowledgment

The authors would like to express their sincere gratitude to the Department of Computer Science at Norfolk State University for providing the academic support and resources necessary to carry out this research. This material is based upon work supported by the National Science Foundation under Grant No. 2424872 and the U.S. Department of Energy's Office of Science (SC) under Award Number DE-SC0025722.

References

[1] Q. Zhu, W. Shen, and S. Liu, “Algorithmic fairness in AI-based recruitment: A review and future research agenda,” AI Soc., vol. 39, no. 3, pp. 781–799, 2024.

[2] M. Soleimani, E. D. P. Oliveira, and P. J. G. Lisboa, “Reducing AI bias in recruitment and selection: An integrative grounded approach,” Int. J. Hum. Resour. Manag., 2023.

[3] A. L. Hunkenschroer and C. Lütge, “Ethics of AI-enabled recruiting and selection: A review and research agenda,” J. Bus. Ethics, 2022.

[4] N. Mehrabi et al., “A survey on bias and fairness in machine learning,” ACM Comput. Surveys, 2021.

[5] T. Pagano, N. Rehman, and I. McLoughlin, “Bias and unfairness in machine learning models: A systematic review,” Big Data Cogn. Comput., vol. 14, no. 2, pp. 45–59, 2023.

[6] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” Science, vol. 356, no: 6334, pp. 183–186, 2017.

[7] Z. Chen, “Ethics and discrimination in artificial intelligence-enabled recruitment practices,” Humanit. Soc. Sci. Commun., vol. 10, 2023.

[8] J. Morley et al., “Ethical guidelines for AI in the context of HR recruitment,” AI Ethics, vol. 2, no. 3, pp. 491–505, 2022.

[9] P. Li and H. Liu, “Achieving fairness at no utility cost via data reweighing with influence,” Proc. Int. Conf. Mach. Learn. (ICML), 2022.

[10] G. Sogancioglu, H. Kaya, and A. Salah, “Using explainability for bias mitigation: A case study for fair recruitment assessment,” Proc. Int. Conf. Multimodal Interaction, 2023.

[11] T. Gebru et al., “Datasheets for datasets,” Commun. ACM, vol. 64, no. 12, pp. 86–92, 2021.

[12] S. Barocas and A. D. Selbst, “Big data’s disparate impact,” Calif. Law Rev., vol. 104, no. 3, pp. 671–732, 2016.

[13] [13] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” Adv. Neural Inf. Process. Syst., 2016.

[14] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” Proc. ITCS, 2017.

[15] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” Proc. FAT, 2018.

[16] A. Chouldechova, “Fair prediction with disparate impact,” Big Data, vol. 5, no. 2, pp. 153–163, 2017.

[17] S. Corbett-Davies and S. Goel, “The measure and mismeasure of fairness,” arXiv preprint arXiv:1808.00023, 2018.

[18] S. A. Friedler et al., “A comparative study of fairness-enhancing interventions,” Proc. FAT, 2019.

[19] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” Knowl. Inf. Syst., 2012.

[20] M. Feldman et al., “Certifying and removing disparate impact,” Proc. KDD, 2015.

[21] M. B. Zafar et al., “Fairness constraints: Mechanisms for fair classification,” Proc. AISTATS, 2017.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining predictions,” Proc. KDD, 2016.

[23] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., 2017.

[24] L. Floridi et al., “AI4People—An ethical framework for a good AI society,” Minds Mach., 2018.

[25] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nat. Mach. Intell., 2019.

[26] M. Raghavan et al., “Mitigating bias in algorithmic hiring,” Proc. FAT, 2020.

[27] A. R. Perera, H. N. P. Nguyen, and T. G. Nguyen, "Bias and Fairness in Artificial Intelligence: A Systematic Review of the Research Landscape and Future Directions," Discover Artif. Intell., vol. 3, no. 1, pp. 1–28, 2023.

[28] K. Wilson, "No Thoughts, Just AI: Human Interaction with Racially-Biased Resume Screening Systems," 2024.

[29] M. Vasconcelos, C. Cardonha, and B. Goncalves, "Modeling Epistemological Principles for Bias Mitigation in AI Systems," Proc. AAAI/ACM Conf. AI, Ethics, Soc., pp. 323–329, 2018.

[30] M. Yoosof, L. Nan, K. Bo, J. Lijffijt, and T. De Bie, "A Challenge-Based Survey of E-recruitment Recommendation Systems," ACM Comput. Surveys, 2024.

[31] F. M. Dena and R. M. Nihar, "Ethical Considerations in AI-Based Recruitment," Proc. IEEE Int. Symp. Technol. Soc. (ISTAS), 2019.

Zero-Disclosure Material Passports for Verifiable Provenance in Multi-Tier Supply Networks

Shivani Dharmavaram
University of the Cumberlands, Kentucky, USA

***Abstract*—Global supply chains produce vast quantities of transactional data, yet most existing traceability systems force companies to choose between disclosing sensitive commercial relationships to a shared infrastructure and relying on mechanisms that do not provide strong privacy guarantees. This study introduces a zero-disclosure material passport framework for verifiable provenance in multi-tier supply networks. The framework represents product history as a Resource-Event-Agent provenance graph and authenticates each event through unlinkable aggregate signatures. Individual participant signatures collapse into a single, fixed-size aggregate proof, so the cryptographic verification object remains 192 bytes in size regardless of the supply chain depth. Selective disclosure is supported through commitment-based predicate proofs that allow regulators and auditors to verify attributes such as origin, certification status, or recycled-content thresholds without learning unrelated commercial information. A Rust prototype was evaluated on an x86-64 workstation, a Raspberry Pi 4, and a Raspberry Pi Zero 2 W. Signing latency remained below 700 ms on the most constrained device, demonstrating feasibility for low-frequency supply chain handover events and offline inspection contexts. The study further provides algorithmic descriptions of credential issuance, passport update, aggregate verification, and selective disclosure; an expanded comparison against blockchain, EDI, BBS+/anonymous-credential, and zk-SNARK-based approaches; and a discussion of the centralized Credential Authority as an explicit trust assumption. Security analysis establishes unforgeability and unlinkability under standard hardness assumptions when the Credential Authority is honest, and participant credentials are issued through the prescribed blinded protocol.**

Keywords*—*Supply chain traceability; material passports; zero-knowledge proofs; aggregate signatures; privacy-preserving authentication; circular economy; provenance verification; selective disclosure

I. Introduction

Tracking a product across a global supply chain appears straightforward in principle, but the operational reality is complex. A finished good may pass through dozens of independent companies spread across several continents, with each organization contributing value while also holding information that competitors, buyers, and even some regulators should not be able to infer. Centralized traceability databases provide a searchable record, but they also concentrate sensitive data such as supplier identities, sourcing patterns, pricing signals, and route dependencies in a single system. Permissioned blockchains distribute records across a consortium, yet their transaction ledgers often remain visible to network members. Even when payloads are encrypted, transaction metadata can expose commercially valuable relationships.

The core problem is that many traceability systems treat verifiability and disclosure as a single objective. Proving that a product history is authentic does not logically require revealing every actor involved in that history. This study separates those objectives by introducing a zero-disclosure material passport: a cryptographic digital twin of a physical product that carries a complete Resource-Event-Agent lineage while revealing only the attributes required by a particular verifier. The term zero-disclosure is used here as a system-level property. It means that the end-to-end architecture is designed to reveal nothing beyond the verifier's authorized disclosure predicate. This is broader than the cryptographic term zero-knowledge [22], which refers to a specific class of proof systems; zero-knowledge proofs are one component of the framework, but zero-disclosure is the design goal of the whole protocol.

The proposed framework combines three ideas. First, BLS aggregate signatures provide a constant-size authentication object for a chain of events. Second, re-randomizable anonymous credentials prevent an observer from linking two credential presentations to the same participant. Third, a Resource-Event-Agent provenance graph represents supply chain activity as resources, events, and role-authorized agents rather than as named companies. A verifier can therefore confirm that the right type of actor performed the right type of operation at the right point in the chain without learning that actor's identity or the surrounding commercial relationship.

The contributions of this study are as follows: First, it defines a zero-disclosure material passport architecture that integrates BLS aggregate signatures, commitment-based selective disclosure, and a Resource-Event-Agent provenance directed acyclic graph (DAG). Second, it provides algorithmic descriptions of the main protocol operations so that the method can be implemented and audited independently. Third, it states the relevant security claims and clarifies the assumptions required for unforgeability and unlinkability, including the fact that the centralized Credential Authority is trusted not to issue fraudulent or traceable credentials. Fourth, it evaluates a working prototype across three hardware platforms, from server-class infrastructure to low-power edge devices. Fifth, it expands the comparative analysis to include conventional EDI, permissioned blockchain, privacy-preserving blockchain, BBS+/anonymous-credential, and zk-SNARK-based approaches. Finally, it discusses the practical meaning of measured latency and verification costs for real industrial deployment.

The remainder of the study is organized as follows: Section II reviews related work and identifies the comparative gap. Section III presents the cryptographic framework and algorithms. Section IV describes the material passport architecture. Section V reports implementation and performance results. Section VI discusses the benchmark results, limitations,

and deployment implications. Section VII presents supply chain integration scenarios. Section VIII concludes with future research directions.

II. Related Work and Comparative Gap Analysis

Research on supply chain traceability draws from cryptographic authentication, privacy-enhancing technologies, and distributed data management. Early traceability systems relied on EDI agreements and centralized databases, which made business-to-business data exchange operationally convenient but did not provide cryptographic privacy for participants whose data were stored or processed by shared infrastructure. Digital signatures improved integrity, but conventional signatures expose the signer and are therefore poorly suited to anonymous or commercially confidential traceability [5]. In terms of the four properties used in this study, such systems may permit offline lookup against a local copy, but they provide neither participant unlinkability nor attribute-level selective disclosure, and they carry no constant-size cryptographic proof object at all.

Digital product passport research has increasingly emphasized privacy and interoperability. Structured reviews of passport systems identify privacy, traceability, lifecycle coverage, and standardization as recurring requirements that existing deployments often address only partially [7], [9], [11]. Blockchain-based provenance systems provide immutable or append-only logs, and systems such as PrivChain aim to protect provenance relationships in blockchain-enabled supply chains [4]. However, blockchain ledgers can still expose participation patterns, and many designs require online verification or proofs that grow as the provenance chain deepens. Hyperledger Fabric provides permissioned channels and strong enterprise integration properties, but channel membership and endorsement metadata can reveal participation patterns unless carefully engineered [19]. Assessed against the four target properties, these designs diverge in opposite ways: PrivChain preserves participant unlinkability and end-to-end lineage but forfeits the constant-size proof and offline verification, since its proofs grow with chain depth and its checks expect a live network, whereas Hyperledger Fabric retains workable record sizes yet surrenders unlinkability because channel and endorsement metadata expose who participated.

More advanced privacy mechanisms have also been proposed. zk-SNARKs allow a verifier to check a statement without learning witness data, and they have been applied to sustainability reporting and digital product passport verification [6]. Their main disadvantages in supply chain settings are operational complexity, circuit maintenance, trusted-setup concerns for some constructions, and limited readability for enterprise teams that must audit the workflow. BBS+ and related anonymous credential systems support selective disclosure and unlinkable presentations, making them highly relevant to privacy-preserving traceability [18]. Their challenge is that they normally address credential presentation rather than the full provenance-chain aggregation problem. A material passport needs both attribute-level disclosure and end-to-end authentication across many supply chain events. Mapped onto the four properties, both families deliver participant unlinkability and attribute-level disclosure, but neither resolves the gap on its own: zk-SNARK constructions can provide succinct verification, yet applying them to a full provenance graph shifts complexity into circuit design, setup choices for some systems, and proving cost, while BBS+ credentials address presentation rather than chain-wide aggregation and so do not supply an end-to-end DAG lineage. This is the combination that the proposed framework is designed to close, as Table I sets out.

Several provenance ontologies have been used in supply chain systems. GS1 EPCIS is widely adopted in retail and pharmaceutical tracking, but it normally records explicit business actors and read-point events. W3C PROV provides a general model for provenance in heterogeneous information systems, but it was not designed around cryptographic unlinkability. The Resource-Event-Agent ontology, originally proposed for accounting systems, models economic activity through resources, events, and agents [23]. That abstraction is useful in this framework because an agent can denote an authorized role rather than a named firm.

The gap that motivates this work is, therefore, not the absence of traceability systems, blockchains, anonymous credentials, or zero-knowledge proof systems. The gap is the absence of a single supply chain passport design that simultaneously provides: 1) a constant-size proof object, 2) offline verification, 3) participant unlinkability, and 4) attribute-level selective disclosure. The reviewed systems tend to satisfy some of these properties but not all. Table I makes this relationship explicit and connects the literature survey to the design requirements used in the proposed framework.

III. Cryptographic Framework Design

The cryptographic engine is built around BLS signatures on elliptic curve groups equipped with a bilinear pairing [5]. The important engineering property is homomorphic aggregation: signatures produced by multiple participants can be combined into a single signature whose size does not depend on the number of signers, a property also studied in efficient multi-signature designs for blockchain settings [21]. This property is essential for deep supply chains because it prevents a passport from growing linearly with the number of enterprises that have handled the product.

Let G_1 and G_2 be the pairing groups of prime order q, let G be a generator, and let $e : G_1 \times G_2 \to G_T$ be a bilinear map. Each participant i holds a secret key sk_i and public key $pk_i = sk_iG$. For an event message m, the participant computes:

$$\sigma_i = sk_i \cdot H(m) \tag{1}$$

where, $H(\cdot)$ hashes the event descriptor to the curve. The aggregate signature is:

$$\sigma = \sum_{i=1}^{N} \sigma_i \tag{2}$$

A verifier checks:

$$e(\sigma, G) = e(H(m), \sum_{i=1}^{N} pk_i) \tag{3}$$

TABLE I. COMPARISON OF THE PROPOSED FRAMEWORK AGAINST PRIOR SYSTEM FAMILIES

System family	Constant proof size	Offline verify	Participant unlinkability	Attribute disclosure	End-to-End DAG lineage	Main limitation
GS1 EDI/EPCIS [11]	No	Yes	No	No	Partial	Mature operational standard, but relies on explicit identifiers and access control.
Hyperledger Fabric [19]	Partial	No	No	Partial	Partial	Permissioned channels help governance, but metadata can reveal participation.
PrivChain [4]	No	No	Yes	Partial	Yes	Protects provenance privacy, but proof and verification costs grow with chain depth.
BBS+/anonymous credentials [16], [18]	Partial	Yes	Yes	Yes	No	Strong credential disclosure primitive, but not a complete provenance aggregation architecture.
zk-SNARK/DPP designs [6]	Partial	Partial	Yes	Yes	Partial	Expressive private verification, but circuit design, setup, and operational complexity remain high.
Proposed framework	Yes	Yes	Yes	Yes	Yes	Relies on a trusted or consortium-operated Credential Authority in the current design.

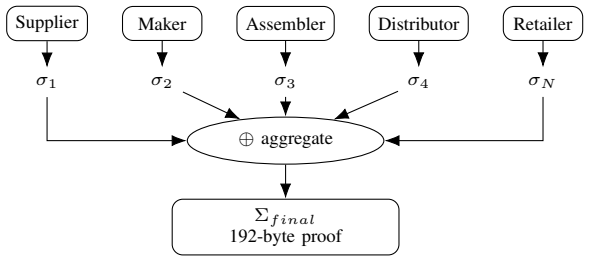


Fig. 1. Signature aggregation. Individual participant signatures are combined into one aggregate authentication proof, whose size is independent of the number of participants.

The result is a fixed-size authentication object even when N grows. In this prototype, the final Event Authentication Code is represented as a 192-byte aggregate proof. Fig. 1 illustrates how individual participant signatures collapse into this single constant-size authentication object.

Participants do not present stable identities during signing. They obtain role-specific credentials from the Credential Authority through blinded issuance and then re-randomize those credentials before each session. Re-randomization changes the public presentation while preserving the algebraic relationships needed for verification. As a result, a verifier can confirm that the signer had the required authorization without learning whether two events were signed by the same organization.

A. Protocol Algorithms

Algorithm 1 to Algorithm 3 describe the protocol in implementation-oriented terms, covering credential issuance, passport update and aggregation, and selective-disclosure verification.

Algorithm 1. Credential Onboarding and Session Randomization

Require: Participant role request R, participant key pair (sk_i, pk_i), public CA key PK_{CA}
Ensure: Re-randomized session credential $cred_i^{(t)}$
1: Participant constructs a blinded request $B(R, pk_i; r)$ using fresh randomness r.
2: Participant sends B to the Credential Authority through an authenticated enrollment channel.
3: Credential Authority checks policy eligibility for role R without learning the unblinded credential presentation.
4: Credential Authority issues a role credential $cred_i$ bound to pk_i and authorized actions.
5: For each signing session t, participant samples fresh randomness ρ_t.
6: Participant computes $cred_i^{(t)} = ReRand(cred_i, \rho_t)$.
7: **return** $cred_i^{(t)}$ for unlinkable event signing.

Algorithm 2. Passport Event Append and Aggregate Authentication

Require: Passport DAG D, event descriptor m, authorized participants $P = \{1, \ldots, N\}$
Ensure: Updated passport D' with aggregate proof Σ_{event}
1: Create an event node E_m containing timestamp, location reference, and committed operating attributes.
2: **for** each participant $i \in P$ **do**
3: Verify that $cred_i^{(t)}$ authorizes the role required by E_m.
4: Compute signature component $\sigma_i = sk_i \cdot H(m)$.
5: **end for**
6: Aggregate signature components $\Sigma_{event} = \sum_{i=1}^{N} \sigma_i$.
7: Link input resource nodes, E_m, role-authorized agent nodes, and output resource nodes in D.
8: Store Σ_{event} and inherited authentication context on the new DAG edges.
9: **return** Updated passport graph D'.

Algorithm 3. Verification and Selective Disclosure

Require: Passport D, aggregate proof Σ, disclosure predicate $P(a)$, public parameters
Ensure: Accept or reject with only authorized attributes disclosed
1: Parse the relevant resource, event, and agent nodes from D.
2: Reconstruct the event message m and authorized public-key aggregate PK_{agg}.
3: Check pairing equation $e(\Sigma, G) = e(H(m), PK_{agg})$.
4: **if** pairing check fails **then**
5: Reject the passport.
6: **end if**
7: For each requested attribute, verify the commitment opening or predicate proof.
8: Run the non-interactive Sigma-protocol verifier for $P(a)$.
9: **if** all predicates and graph links verify **then**

10: Accept and reveal only the attributes allowed by $P(a)$.
11: **else**
12: Reject.
13: **end if**

B. Selective Disclosure Construction

Each resource attribute a is stored as a commitment rather than plaintext. The prototype uses Pedersen-style commitments [17] of the form:

$$C = g^a h^r \tag{4}$$

where, g and h are public group generators and r is a uniformly sampled blinding factor. To prove a predicate $P(a)$, such as recycled content exceeding a threshold, the passport holder constructs a non-interactive Sigma-protocol proof of knowledge showing that the committed value satisfies both $C = g^a h^r$ and $P(a)$ without revealing a or r directly. The Fiat-Shamir transform makes the proof non-interactive, which is important for asynchronous supply chain settings and offline verification.

C. Threat Model and Trust Assumptions

The adversary is modeled as a probabilistic polynomial-time algorithm that may observe all public credential presentations, aggregate proofs, and protocol transcripts. It may corrupt any subset of ordinary supply chain participants and obtain their local signing keys and credentials. It may also attempt replay, double-signing, and graph-splicing attacks by recombining valid events into invalid lineages.

A central trust assumption is that the Credential Authority is not corrupted. This assumption is explicit because the Credential Authority issues the role credentials that make unlinkable authorization possible. If the Credential Authority were malicious, it could issue fraudulent credentials, refuse legitimate credentials, or embed traceable structure in credential issuance. The current design therefore provides security under an honest-but-policy-enforcing Credential Authority model. Threshold credential issuance and multi-authority issuance are treated as future work because they remove this institutional single point of trust but require additional coordination and governance.

The Credential Authority assumption does not weaken the aggregate-signature invariance property, but it does define the boundary of the privacy and authorization guarantees. In practical deployments, the Credential Authority should be operated by a regulator, standards body, industry consortium, or threshold group rather than by a single supply chain participant.

D. Security Claims

Security Claim 1 (Unforgeability): Assuming co-Diffie-Hellman hardness in the selected pairing group and proper proof-of-possession or CA-mediated key registration, an adversary that does not hold a valid credential for an authorized role cannot produce an aggregate signature component that passes the verification equation for a new event, except with negligible probability.

Reduction outline. Suppose an adversary produces a valid aggregate proof for an event requiring at least one uncorrupted authorized signer whose signature component was never issued for that message. The reduction embeds a co-CDH challenge in that target signer's public key. Signatures for corrupted or honestly simulated participants are known to the reduction and can be algebraically subtracted from the forged aggregate because aggregation is linear. What remains is a valid BLS signature on the challenge public key for a message that was not queried, yielding a solution to the underlying co-CDH/BLS forgery game. Rogue-key attacks are excluded by the CA-issued credential binding and proof-of-possession requirement during onboarding. Thus, a successful aggregate forgery implies a successful underlying BLS forgery, contradicting the hardness assumption [5].

Security Claim 2 (Unlinkability): Assuming decisional Diffie-Hellman hardness for the credential presentation group, an adversary observing two valid credential presentations cannot distinguish whether they were generated by the same participant or by two independent participants, except with negligible advantage.

Reduction outline. Each session presentation applies fresh randomization to the participant credential. The resulting group elements are distributed as randomized commitments preserving authorization relations but hiding the stable credential representation. If an adversary could link two such presentations with non-negligible advantage, a simulator could use that adversary to distinguish a valid Diffie-Hellman tuple from a random tuple by embedding the challenge tuple in the credential randomization. This contradicts the DDH assumption and follows the unlinkability logic of re-randomizable anonymous credentials [16], [18].

IV. Material Passport Architecture

A material passport is a signed, portable graph document that encodes what a verifier needs to confirm a product's lineage: what resource was handled, what event occurred, what role-authorized agent performed the action, and how that event relates to upstream and downstream resources. The graph is a DAG in which nodes and edges carry authentication context. This prevents a downstream verifier from accepting a product state unless its lineage links back through authenticated transformations. Fig. 2 summarizes the architecture connecting credential issuance, event signing, provenance-graph construction, and selective-disclosure verification.

A. Resource Nodes

A resource node captures a material, component, or product at a particular point in its lifecycle. The technical attributes it records, such as composition, grade, batch quantity, and dimensions, are held as commitments rather than as plain values. Crucially, the node keeps no clear record of the supplier, manufacturer, or handler involved. Because each attribute is committed separately, a verifier can later confirm any single claim without the holder having to reveal the complete attribute vector.

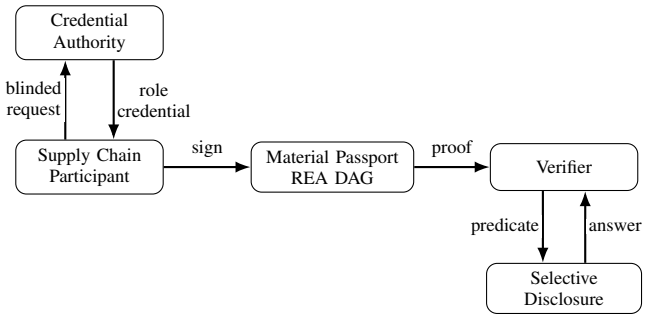


Fig. 2. Zero-disclosure material passport architecture. Blinded credential issuance, unlinkable event signing, graph-based provenance, and selective disclosure are combined in one verification workflow.

B. Event Nodes

An event node describes something that happens to a resource, whether a transformation, a measurement, an inspection, or a transfer of custody. Alongside timestamp metadata, a location reference, and the relevant operating parameters, each event carries an aggregate proof confirming that the required roles authorized it. Where an operation is complex, it can be broken down hierarchically: a sorting event, for instance, may branch by material class and then split further into sub-branches for polymer or alloy type. This layering gives auditors fine-grained insight while still allowing the passport to withhold details that a particular verifier does not need to see.

C. Agent Nodes

Agent nodes stand for the role-authorized actors in the chain. Rather than recording a company name or a tax identifier, an agent node holds an unlinkable credential commitment together with a role certificate. A verifier can therefore establish that a customs broker, recycler, certified manufacturer, or authorized inspector took part in an event, yet learn nothing about which specific organization sits behind that role.

D. Edge Authentication and Lineage Propagation

Edges in the DAG express the parent-child relationships between resources and events. When a transformation turns input resources into output resources, the resulting output node is linked both to those inputs and to the governing event node. Since the output node inherits its authentication context from the event proof, a downstream verifier can trace the chain backward without ever contacting the participants upstream. This property is what makes offline inspection workable in ports, warehouses, and field environments.

E. Physical-Digital Coupling

Each digital passport is bound to a physical item or container by a cryptographic seed embedded in a hardware identifier such as an RFID tag, a QR code, or a serialized label. Scanning that seed triggers a verification query, which runs either against a network backend or against a locally cached copy of the passport. For bulk materials, the seed can be fixed to the sealed container instead of to individual units, which preserves custody integrity at the shipment level.

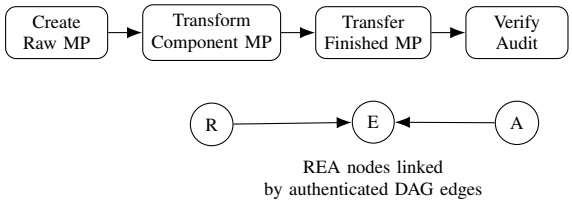


Fig. 3. Material passport lifecycle. Resource, Event, and Agent nodes are extended, as the product is created, transformed, transferred, and verified.

TABLE II. OPERATION TIMINGS AND OUTPUT SIZES

	Time (ms)			Size (B)		
Operation	X64	RPi4	RPi0	X64	RPi4	RPi0
Credential	13.7	37.3	249.2	96	96	96
Session	18.8	51.6	276.0	128	128	128
Signing	73.0	145.6	660.3	64	64	64
Verify	40.5	104.9	575.3	0	0	0
Aggregate	21.9	57.9	346.9	192	192	192
Disclosure	39.1	88.3	516.5	48–256	48–256	48–256

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

The prototype is implemented in Rust to obtain memory safety and predictable runtime behavior. Cryptographic execution is sandboxed so that signing state and credential material are isolated across operations. The provenance graph is serialized in JSON-LD, allowing semantic data and cryptographic proofs to be processed by standard graph tooling while still supporting compact verification artifacts.

System setup begins when the Credential Authority generates public parameters for the BLS12-381 curve and publishes the system verification key. Participants generate local signing keys and request role credentials through the blinded issuance protocol in Algorithm 1. At runtime, event nodes are appended using Algorithm 2; verification and selective disclosure follow Algorithm 3.

Fig. 3 summarizes the implemented passport lifecycle across creation, transformation, transfer, and verification, showing how credentials, event signatures, REA graph updates, and selective-disclosure checks are connected during execution.

The evaluation uses three hardware classes: an x86-64 workstation representing cloud or enterprise infrastructure; a Raspberry Pi 4 representing handheld or gateway devices; and a Raspberry Pi Zero 2 W representing constrained edge hardware. Table II reports operation timings and output sizes. The most resource-constrained signing operation completes in 660.3 ms, and verification completes in 575.3 ms. These values are appropriate for supply chain handovers, inspections, and batch-level traceability events, which occur at human or process timescales rather than at high-frequency sensor rates.

Table III shows scalability as participant count and stage count increase. Signing and aggregation work grow with the number of participants because each authorized role must contribute a signature component. However, the proof object carried by the passport remains fixed. The total passport size grows primarily with semantic provenance records and disclosed metadata, not with the aggregate cryptographic object.

TABLE III. SCALABILITY ANALYSIS WITH INCREASING PARTICIPANTS

Participants	Stages	Sign (ms)	Verify (ms)	Size (KB)
5	3	365	122	1.2
10	5	730	258	2.0
50	8	3,650	797	3.2
100	12	7,300	1,943	4.8
500	20	36,500	13,974	8.0
1,000	25	73,000	35,230	10.0

High-throughput distribution centers can use batch verification to amortize pairing computations across many concurrent checks. Handheld devices can use cached public parameters and pre-computed lookup tables to support offline operation. Full audit tools can recursively verify every graph edge and aggregate proof in a passport's history, which is slower but appropriate for forensic or regulatory review.

VI. DISCUSSION

The benchmark results should be interpreted in relation to the research question introduced in Section I: Can supply chain verifiability be achieved without exposing commercially sensitive participant relationships? The results indicate that the answer is yes for low-frequency provenance events such as custody transfer, batch certification, transformation completion, and inspection. The framework's central invariance property is the fixed-size aggregate proof. Even when the number of enterprises increases, the authentication object remains 192 bytes. This is the most important engineering result because it prevents the proof from becoming unusable as the supply chain deepens.

Latency matters differently across deployment tiers. On server infrastructure, signing and verification are fast enough for interactive enterprise systems. On the Raspberry Pi 4, operations remain compatible with handheld inspection and gateway-class devices. On the Raspberry Pi Zero 2 W, signing below 700 ms and verification below 600 ms are not appropriate for high-frequency sensor streams, but they are appropriate for supply chain events that occur at discrete operational checkpoints. This distinction is important: the framework is not intended to sign every raw temperature sample. Instead, it is intended to sign meaningful provenance events, such as a certified temperature excursion report or a completed cold-chain handover.

The verification cost on constrained hardware also has a deployment implication. Offline verification is feasible, but devices should pre-cache public parameters, role-policy metadata, and recent revocation information during synchronization windows. This design is appropriate for ports, warehouses, and field inspection sites where connectivity can be intermittent. When connectivity is available, online verification provides the freshest policy and revocation state. When connectivity is not available, cached verification preserves operational continuity while limiting the freshness of revocation checks.

The expanded comparison in Table I clarifies the position of the framework. BBS+ credentials and zk-SNARKs are not competitors to be dismissed; they are important privacy primitives. The proposed contribution is the integration of privacy-preserving credential presentation, fixed-size event aggregation, and graph-structured provenance into a passport architecture that supports offline verification. In this sense, the framework is an applied systems design rather than a claim that any individual primitive is new.

The most important current limitation is the centralized Credential Authority. The threat model assumes that the Credential Authority does not issue fraudulent credentials and does not embed traceable structure into credentials. This assumption is reasonable for a prototype and for some regulated settings, but it is not ideal for global industrial deployment. A malicious or compromised Credential Authority could damage both trust and privacy. Future deployments should therefore use threshold credential issuance, multi-authority governance, external audits, and transparent credential policies. Another limitation is that high-frequency sensor attestation would require batching, hardware acceleration, or a separate streaming attestation layer.

VII. SUPPLY CHAIN INTEGRATION CASE STUDIES

The framework applies to several supply chain contexts in which auditability and confidentiality are simultaneously required.

A. Pharmaceutical Supply Chains

Drug manufacturing requires evidence of ingredient origin, manufacturing conditions, packaging integrity, and handover history. A zero-disclosure passport can authenticate batch-level movement from active pharmaceutical ingredient synthesis through formulation and distribution while hiding which contract manufacturers or logistics providers participated in a particular route [3]. Regulators can verify compliance, while commercial sourcing arrangements remain private.

B. Electronics and Conflict Minerals

Electronics supply chains often span many tiers, and manufacturers may need to prove that minerals were not sourced from prohibited regions. A zero-disclosure passport can prove that each upstream tier satisfied the required certification predicate without naming every supplier. This supports conflict-mineral auditability while preserving sourcing strategy and supplier confidentiality [10].

C. Circular Economy and Material Reuse

Circular economy certification depends on proving that reused or recycled materials came from the claimed source and passed through valid transformations. The graph structure naturally supports split and merge operations: one recycled material batch may divide into multiple product streams, and multiple streams may later combine into one component. Each branch inherits cryptographic provenance from the authenticated DAG [2].

D. Food Safety and Farm-to-Fork Traceability

Farm-to-fork systems require certification of origin, pesticide use, organic status, and cold-chain integrity. Selective disclosure allows investigators to identify an affected batch during a contamination event without exposing unrelated producers or unaffected shipment data. This enables narrower recalls and reduces collateral reputational harm [15].

E. Construction Materials

Construction projects increasingly require auditable evidence of embodied carbon, recycled content, and material origin. Material passports can preserve the provenance of cement, aggregate, steel, and admixtures from source to installation. This is useful for green-building certification, circular reuse, and forensic investigation after structural failures [1].

F. Aerospace and Defense

Aerospace and defense supply chains require part-level traceability but also involve sensitive supplier relationships. A passport can prove that a component passed through authorized custody and inspection steps while concealing supplier identities when disclosure is not legally required [13]. This reduces counterfeit-part risk without unnecessarily exposing the supply network.

G. Cross-Border Trade and Customs Compliance

Customs authorities need proof of origin, export-license status, and regulatory compliance, but they do not need full commercial routing or pricing data. Selective disclosure can provide a proof package tailored to customs requirements. Because verification can operate offline, border clearance is not blocked when foreign systems are unavailable [6].

H. Precision Recall Management

Product recalls are often broader than necessary because manufacturers cannot prove which units are affected. A full provenance DAG allows a defective batch, production line, or time window to be identified cryptographically. Downstream verifiers can determine whether a unit belongs to the affected lineage without seeing unrelated production data [12].

I. Industrial IoT Integration

Industrial sensors can provide temperature, humidity, vibration, and location readings that are relevant to compliance. The framework should not sign every raw reading on constrained hardware; instead, it can sign summarized, policy-relevant events such as excursion reports, tamper alerts, or inspection confirmations. Sensor-based passport designs demonstrate the practical value of binding environmental monitoring to product records [14].

J. Regulatory Reporting Automation

Regulatory reporting often requires evidence that already exists in operational systems but is difficult to assemble and verify. A material passport can serve as the authoritative evidence graph, from which jurisdiction-specific reports are generated using selective disclosure. Each authority receives only the attestations relevant to its legal scope [8].

VIII. Conclusion and Future Directions

This study has presented a zero-disclosure material passport framework for verifiable provenance in multi-tier supply networks. The design brings together four building blocks: BLS aggregate signatures, unlinkable credential presentations, commitment-based selective disclosure, and a Resource-Event-Agent provenance DAG. Its central engineering result is a constant-size aggregate proof. However many enterprises have handled a product, the authentication object stays at 192 bytes, which is what keeps the scheme practical as supply chains grow deeper. Prototype benchmarks on an x86-64 workstation, a Raspberry Pi 4, and a Raspberry Pi Zero 2 W confirm that the approach is workable for low-frequency provenance events and for offline inspection in the field.

The framework's position relative to prior work is equally clear. Existing EDI, blockchain, BBS+/anonymous-credential, and zk-SNARK-based approaches each contribute valuable capabilities, yet none of them delivers constant-size proofs, offline verification, participant unlinkability, attribute-level disclosure, and end-to-end graph lineage at the same time. The contribution here is to combine those properties within a single passport workflow while stating the centralized Credential Authority openly as a trust assumption rather than leaving it implicit.

Several directions remain open. The most pressing is the threshold credential issuance, which would spread authority across several independent parties so that no single institution controls who may sign. Post-quantum readiness is a second priority, since pairing-based signatures will not withstand a sufficiently capable quantum adversary; lattice-based and hash-based schemes such as CRYSTALS-Dilithium [20] are natural candidates, provided they can preserve the aggregation and unlinkability the framework depends on. Beyond cryptography, further work could address high-frequency sensor attestation, the automatic extraction of structured event metadata from shipping documents, and independent security audits of open-source reference implementations. As digital product passport regulations continue to widen, zero-disclosure material passports offer a credible way to make global supply chains more accountable without forcing companies to expose the commercial relationships that keep them competitive.

References

[1] A. Vahidi, A. T. Gebremariam, F. Di Maio, K. Meister, T. Koulaeian, and P. Rem, "RFID-based material passport system in a recycled concrete circular chain," *Journal of Cleaner Production*, vol. 442, p. 140973, 2024.

[2] S. Wilson, K. Adu-Duodu, Y. Li, R. Sham, M. Almubarak, Y. Wang, E. Solaiman, C. Perera, R. Ranjan, and O. Rana, "Blockchain-enabled provenance tracking for sustainable material reuse in construction supply chains," *Future Internet*, vol. 16, no. 4, p. 135, 2024.

[3] A. Padma and M. Ramaiah, "Blockchain based solution for secure information sharing in pharma supply chain management," *Heliyon*, vol. 10, no. 22, p. e40273, Nov. 2024.

[4] S. Malik, V. Dedeoglu, S. S. Kanhere, and R. Jurdak, "PrivChain: Provenance and privacy preservation in blockchain enabled supply chains," in *Proc. 2022 IEEE Int. Conf. Blockchain*, Espoo, Finland, Aug. 2022, pp. 157–166.

[5] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the Weil pairing," *Journal of Cryptology*, vol. 17, no. 4, pp. 297–319, 2004.

[6] C. Udokwu and S. Crass, "Zero-knowledge proof extensions for digital product passports in sustainability claims reporting and verifications," *Electronics*, vol. 15, no. 4, p. 745, Feb. 2026.

[7] A. KC, S. Senaratne, S. Perera, and S. Nanayakkara, "Review of current digital technologies for material passports to enhance circularity towards net zero," *Built Environment Project and Asset Management*, ahead-of-print, 2024.

[8] D. Tortola, C. Felicioli, A. Canciani, and F. Severino, "Authenticated data visualization for hybrid blockchain-based digital product passports," *Computer Communications*, vol. 235, pp. 1–12, 2025.

[9] C. Udokwu, V. Patrick, A. Norta, and D. Draheim, "State of research on digital product passports: A structured literature review," *IEEE Access*, vol. 11, pp. 114059–114085, 2023.

[10] A. Canciani, C. Felicioli, F. Severino, and D. Tortola, "Enhancing supply chain transparency through blockchain product passports," in *Proc. 2024 IEEE Int. Conf. Pervasive Computing and Communications Workshops*, Biarritz, France, Mar. 2024, pp. 751–756.

[11] M. Jansen, T. Meisen, C. Plociennik, H. Berg, A. Pomp, and W. Windholz, "Stop guessing in the dark: Identified requirements for digital product passport systems," *Systems*, vol. 11, no. 3, p. 123, Feb. 2023.

[12] K. Berger, M. Rusch, A. Pohlmann, M. Popowicz, B. C. Geiger, H. Gursch, J.-P. Schoeggl, and R. J. Baumgartner, "Confidentiality-preserving data exchange to enable sustainable product management via digital product passports: A conceptualization," *Procedia CIRP*, vol. 116, pp. 354–359, 2023.

[13] S. De Diego and I. Gutierrez-Aguero, "Decentralized digital product passport building blocks for enhancing supply chain sovereignty and circular economy practices," *IEEE Access*, vol. 13, pp. 1–15, 2025.

[14] A. Pracucci and M. Giovanardi, "Design of a sensor-based digital product passport for low-tech manufacturing: Traceability and environmental monitoring in bio-block production," *Sensors*, vol. 25, no. 18, p. 5653, Sep. 2025.

[15] C. Lopes and J. Barata, "Digital product passport: A review and research agenda," *Procedia Computer Science*, vol. 246, pp. 981–990, 2024.

[16] J. Camenisch and A. Lysyanskaya, "An efficient system for non-transferable anonymous credentials with optional anonymity revocation," in *Advances in Cryptology – EUROCRYPT 2001*, Lecture Notes in Computer Science, vol. 2045. Berlin, Germany: Springer, 2001, pp. 93–118.

[17] T. P. Pedersen, "Non-interactive and information-theoretic secure verifiable secret sharing," in *Advances in Cryptology – CRYPTO 1991*, Lecture Notes in Computer Science, vol. 576. Berlin, Germany: Springer, 1992, pp. 129–140.

[18] S. Tessaro and C. Zhu, "Revisiting BBS signatures," in *Advances in Cryptology – EUROCRYPT 2023*, Lecture Notes in Computer Science, vol. 14008. Cham, Switzerland: Springer, 2023, pp. 691–721.

[19] E. Androulaki *et al.*, "Hyperledger Fabric: A distributed operating system for permissioned blockchains," in *Proc. 13th EuroSys Conf.*, Porto, Portugal, Apr. 2018, pp. 30:1–30:15.

[20] L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, P. Schwabe, G. Seiler, and D. Stehle, "CRYSTALS-Dilithium: A lattice-based digital signature scheme," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2018, no. 1, pp. 238–268, 2018.

[21] M. Kara, A. Laouid, and M. Hammoudeh, "An efficient multi-signature scheme for blockchain," Cryptology ePrint Archive, Paper 2023/078, 2023. [Online]. Available: https://eprint.iacr.org/2023/078

[22] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof systems," *SIAM Journal on Computing*, vol. 18, no. 1, pp. 186–208, Feb. 1989.

[23] W. E. McCarthy, "The REA accounting model: A generalized framework for accounting systems in a shared data environment," *The Accounting Review*, vol. 57, no. 3, pp. 554–578, Jul. 1982.

Exploring the Structure Resulting from Unstructured Neural Network Pruning

Jamil Gafur, Max Milkert, Kevin Patrick Griffin, Nicholas T. Wimer, Charles Tripp, Steve Goddard
Computational Science Center, NLR, USA[1,2,3,4]
Computer Science Dept, Univ. of Iowa, USA[1,6]
Computer Science Dept, Vanderbilt Univ., USA[2]
Zazzle, USA[5]

***Abstract*—Iterative Magnitude Pruning (IMP) is a widely used technique for compressing neural networks by progressively removing low-magnitude weights while maintaining predictive accuracy. Despite its widespread application and simplicity, the underlying reasons for its effectiveness remain underexplored. In this work, the pruning dynamics and emergent structural traits of IMP are empirically examined across three benchmark convolutional architectures (ResNet20, Vgg16, and RegNetX) and evaluated on the CIFAR10 and Tiny ImageNet datasets. A comprehensive analysis reveals that IMP preferentially removes weights from deeper layers, preserving early feature extractors until a critical network sparsity threshold of 96% is reached. Up to this 96% threshold, test accuracy remains remarkably stable across all evaluated networks before experiencing significant degradation. Quantitative evidence is provided showing that later stages are pruned more heavily, taking advantage of the fact that deeper layers contain more low-magnitude weights. Furthermore, neuron-level investigations reveal that IMP does not produce any completely pruned (100% sparse) neurons, even at extreme sparsity levels. Sensitivity analysis via neuron zeroing demonstrates that individual neurons maintain a stable range of functional importance rather than narrowing as pruning progresses. Finally, activation similarity metrics indicate that feature representations are preserved throughout the pruning process, keeping pruned and unpruned networks in close representational alignment. These findings highlight the surprising degree of structure generated by unstructured pruning, offering new insights into network compression and potential pathways for hardware-efficient sparse matrix adaptations.**

***Keywords*—Machine learning; convolutional neural networks; multi-layer perceptrons; artificial intelligence**

I. Introduction

As neural networks grow in both complexity and size, so too does the need for additional computing capacity and energy. This creates a challenge in hosting neural network-based applications, particularly on resource-constrained devices. A common solution to this challenge is to prune (i.e., remove) excess parameters or components of the neural networks, thereby reducing the memory and/or energy requirements of the application while retaining its accuracy [1], [2], [3], [4].

A key challenge in pruning is identifying parameters that significantly contribute to model performance. The process determines which parameters can be removed without substantially affecting accuracy. Denil et al. [5] show that a neural network can be effectively reconstructed using only a subset of its parameters. This result demonstrates the potential for model compression while preserving performance, highlighting the importance of careful parameter selection in pruning. Neural network pruning is generally classified into two categories: unstructured and structured pruning, each with distinct trade-offs in sparsity, computational cost, and resulting hardware compatibility of the pruned network.

Unstructured pruning achieves higher compression rates and retains accuracy more effectively at higher sparsity levels. However, Han et al. [6] state that the resulting irregular sparsity patterns are difficult to exploit on modern hardware. In contrast, structured pruning aligns better with hardware accelerators but may significantly degrade accuracy if not applied carefully [7], [8].

Among unstructured methods, IMP [9] has gained popularity for reducing the size and computational demand of neural networks. IMP works by repeatedly training the network and then removing the surviving weights with the smallest magnitude. It then "rewinds" the values of the remaining weights back to a point near initialization and repeats the process (see Fig. 1). Despite its simplicity, IMP achieves comparable or better accuracy retention at high compression ratios when compared to more sophisticated techniques [10]. As identified in Han et al. [6], it is common to use a sparse matrix to represent the parameters of the pruned matrix, which results in less efficient use of hardware accelerators (diminishing the possible benefits of pruning).

This work empirically investigates whether IMP indeed yields an unstructured neural network, or whether sufficient structure is present to achieve the desired memory and energy savings in the pruned network. The pruning experiments (see Fig. 2) reveal the following structure and processing traits:

- Later layers are pruned much more than early layers at all levels of sparsity (until accuracy falls below acceptable levels).
- No neurons are completely removed. That is, no neurons have all of their input weights set to zero.
- Stable neuron-level sensitivity distributions. That is, no single neuron can be removed without impacting accuracy.
- Feature representations (as measured via activation patterns) are preserved throughout the pruning process.

Thus, the results are mixed. We have discovered that IMP does create a surprising amount of structure in the pruned

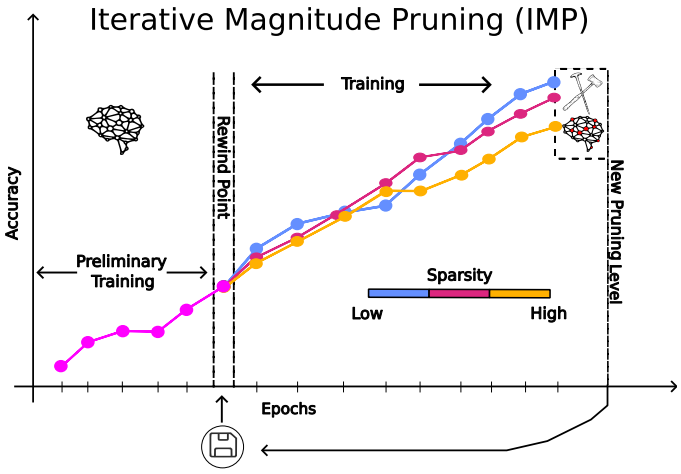


Fig. 1. An illustration of the IMP implementation, based on the approach described in [9], is as follows: Initially, an unpruned "Baseline" model undergoes preliminary training for a specified number of epochs. At the end of this phase, the model is saved as a "rewind point". From there, the network is further trained for a set number of epochs and saved again. After saving, the model is pruned to a new sparsity level, and its weights and biases are reset to those from the rewind point. The pruned weights are then set to zero and made untrainable. This process is repeated until the desired pruning level is achieved.

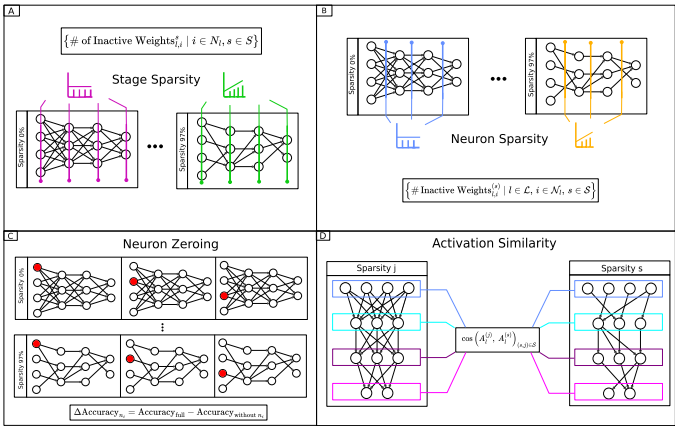


Fig. 2. Overview of the experimental procedures used to analyze sparse neural networks: (A) Stage Sparsity is a metric that calculates the fraction of pruned weights for each neuron across different stages and sparsity levels, providing insight into how pruning is distributed across the network's architecture. (B) Neuron Sparsity measures the percentage of pruned weights for each neuron within a specific layer at various sparsity levels, offering a characterization of the pruning distribution across stages. (C) Neuron Zeroing, where each neuron in a prunable layer is independently ablated by zeroing its incoming weights and bias. The resulting drop in accuracy quantifies the functional importance of each neuron. (D) Activation Similarity, where neuron activations are recorded on a held-out test set at different sparsity levels. These activations are compared using cosine similarity to evaluate how internal representations evolve with pruning.

model, but it is not yet clear how to take advantage of the resulting structure, though we do offer some insights for future exploration.

This study is organized as follows: Section II reviews prior work on pruning and internal network analysis. Section III describes the pruning algorithm, datasets, architectures, and evaluation methods. Section IV presents the experimental findings, and Section V draws conclusions and discusses future research directions.

II. Background and Related Work

Unstructured pruning of convolutional neural networks (CNNs) targets individual weights or connections between neurons for removal, with IMP being among the most widely studied of these techniques [2], [10]. The goal is to remove the "least important" weights, while maintaining the model's original level of accuracy. For IMP, this importance is based on the relative magnitude of the weight. IMP, as shown in Fig. 1, starts by pretraining a network, pruning a fraction of low-magnitude weights from layers with trainable parameters (i.e., pruneable layers), and resetting the remaining weights to an earlier training point, often referred to as the "rewind point". Training is then resumed from this rewound state for the new sparsity level. This process is repeated until the required sparsity level is achieved or a significant degradation in accuracy occurs, at which point the process ends.

Based on this concept, the Lottery Ticket Hypothesis (LTH), by Frankle and Carbin [9], posits that a randomly initialized, sparse sub-network of the original CNN exists. When trained independently from the beginning, it can achieve the performance of the original network with a fraction of the weights. Such sparse sub-networks are called "winning tickets". In this work, the experiments follow the methodology outlined by Frankle and Carbin in [11] for IMP of CNNs to identify winning tickets. Thus, the network is pretrained to allow its parameters to enter an "optimal" loss basin within a few epochs of training. By doing so, the network weights to encode useful information and enter a near-optimal loss basin prior to pruning. This methodology increases the likelihood of finding winning tickets during the IMP process.

Most work on pruning in the literature details the percentage of weights in the entire network that have been pruned. This is commonly referred to as network sparsity. This sparsity is measured across all "stages" in a network, where a stage is a single repeated set of consecutive layers in a network. Section IV-A explores how stage sparsity differs from network sparsity at various pruning levels. Fig. 2(A) illustrates this concept.

Han et al. [2] used a slightly different pruning methodology from IMP. Whereas IMP prunes using weights alone, their methodology first learns the most important weights *and* connections and then prunes "low-weight connections". In their work, a mask was used for each weight tensor to disregard pruned parameters during inference. Thus, their approach did not reduce memory usage. However, their pruning strategy resulted in neurons in which all input weights for a neuron, *and* the weights connected to the neuron's output, were zero. These neurons with "zero input connections or zero output connections" were then pruned from the network with a matching tensor mask.

In this study, the impact of IMP on neuron sparsity (i.e., the percentage of zeroed-out input weights to a given neuron) is examined. Unexpectedly, the results found that IMP yielded no 100% sparse neurons. Fig. 2(B) illustrates conceptually how neuron sparsity increases as the network is pruned. Section IV-B explores this in detail, including how close

neurons were to having all of their input weights be zero across different sparsity levels.

A growing line of research in pruning involves sensitivity analysis, where individual neurons or weights are zeroed out, and the resulting drop in accuracy is measured to estimate their importance to the overall model [3], [12]. Sensitivity analysis reveal that not all neurons contribute equally to a network's prediction; some neurons play a disproportionately important role in determining a network's performance. Section IV-C uses this approach to evaluate the impact of removing each neuron in a layer on the accuracy of the model, as illustrated in Fig. 2(C).

Pieterse et al. [13] implemented structured pruning effects on activation patterns across layers. They used cosine similarity [Eq. (1)] to "determine the importance of a connection based on the activations of the two neurons of that connection". Cosine similarity compares the geometric closeness of activation vectors between pruned and unpruned neurons, providing insights into the preservation of feature representations. Here, cosine similarity is defined as the dot product of two vectors, A and B, normalized by the product of their Euclidean norms.

$$\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

In Section IV-D, the experiments explore unstructured pruning effects on activation patterns across layers using a cosine similarity function. Finally, as shown in Fig. 2(D), pairwise activation cosine similarity matrices are computed for each prunable layer at different pruning levels.

Our use of activation similarity is related to the work of Roychowdhury et al. [14], who used cosine similarity between neuron weight vectors to quantify redundancy. However, such weight-based similarity may not capture functional similarity if upstream activations differ. In contrast, cosine similarity is computed directly on neuron outputs, providing a more complete picture of functional overlap.

III. Methodology

To evaluate the impact of IMP on CNN architectures, the experiment focused on a representative subset of benchmark networks: ResNet20 [15], RegNetX [16], and Vgg16 [17]. As shown in Fig. 3, the networks were trained and pruned using the CIFAR10 [18] and Tiny ImageNet datasets [19]. RegNetX, Vgg16TinyImageNet, Vgg16CIFAR10, and ResNet20 consist of approximately 11.6M, 40.7M, 33.6M, and 272k total number of parameters, respectively.

The experiment focuses on image classification models that can be pruned to fit on resource-constrained devices. CNNs are studied because they have been shown to be significantly compressed with IMP. These models were selected because they are widely used CNN architectures with distinct design philosophies (i.e., residual connections, block scaling, and deep sequential convolutional stacks). Together with the two aforementioned datasets, these models cover both the lightweight and larger scale CNN regimes, making them represent a broad class of widely used networks.

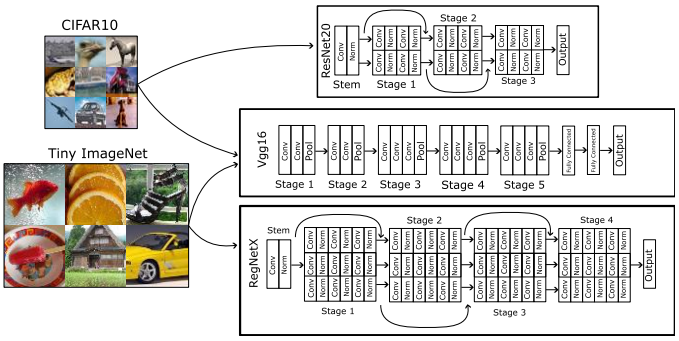


Fig. 3. An illustrative example of the evaluated CNN architectures. From top to bottom: ResNet20, Vgg16, and RegNetX, each shown in their CIFAR10 and Tiny ImageNet configurations. Architectures for Tiny ImageNet are scaled to account for the higher resolution and increased input complexity. Each multi-column layer in ResNet20 and RegNetX is referred to as a "block". This visualization highlights the relative depth, layer types, and structural differences across networks.

Each network underwent 1, 5, or 10 epochs of preliminary training, followed by an additional 30, 80, or 150 training epochs for each network sparsity level. After completing all training for a sparsity level, the network's weights and biases are saved for future analysis. A total of 18 sparsity levels were performed, achieving up to 99% sparsity, following an exponential decay schedule that progressively removed network parameters. This setup followed the work of [9] pruning only the weights of the network. All networks were trained with a batch size of 2048 to maximize hardware utilization. Optimization was carried out using stochastic gradient descent (SGD) with momentum 0.9 and weight decay set to 5×10^{-4} (or 1×10^{-5} for RegNetX). Learning rate schedules varied by architecture: Vgg16 and ResNet20 used a stage decay schedule, while RegNetX employed cosine annealing.

Models Vgg16 and ResNet20 used stage-wise decay schedules, while RegNetX employed cosine annealing. Vgg16's models reduced the learning rate by a factor of 10 every 30 epochs following a fixed interval step decay. On the other hand, ResNet20 had a constant learning rate for the first 50% of training for each pruning stage, then dropped by a factor of 10 until 75%, and was further reduced by another factor of 10 for the remaining epochs in the stage. Finally, RegNetX used cosine annealing during the pretraining phase, gradually reducing the learning rate from an initial value of 0.1 to a minimum of 10^{-6}.

The accuracy of the pruned networks was measured to validate the pruning process. Fig. 4 plots the accuracy of each of the four trained networks against their corresponding sparsity level. The accuracy of all four models is consistent with reported results in the literature for the corresponding model and dataset. He et al. [15] report approximately 91% accuracy on the CIFAR10 dataset using ResNet-20. Similarly, Giuste et al. [20] achieved 93% accuracy with VGG16 on CIFAR10 and approximately 60% accuracy on the ImageNet dataset using the same architecture. Radosavovic et al. [21] report a higher performance, reaching 70% accuracy on ImageNet. Given that the experiments were conducted on the Tiny ImageNet dataset, there was an expected drop in performance. This is because Tiny ImageNet has a lower resolution and

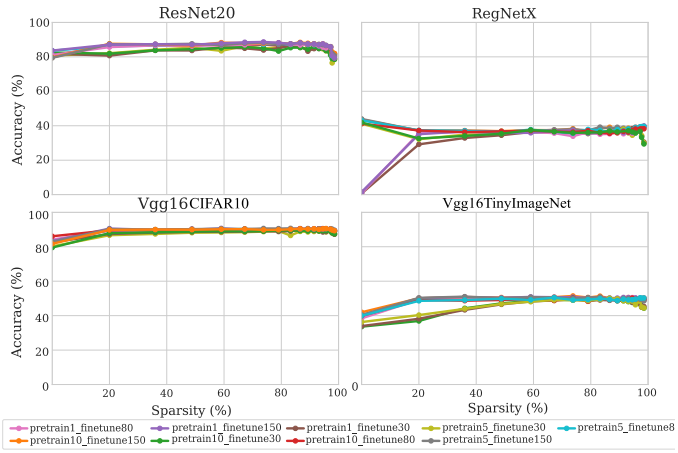


Fig. 4. Accuracy vs. Sparsity. Test accuracy as a function of global weight sparsity for four image classification networks: ResNet20, RegNetX, Vgg16-CIFAR10, and Vgg16-TinyImageNet. Each curve represents a different fine-tuning schedule, applied over 18 pruning iterations.

fewer data points than the full model. TinyImageNet contains 100,000 images across 200 classes with an image size of 64x64. ImageNet is 13 million images across 21841 classes with a resolution size of 256x256 [19].

Fig. 4 also illustrates the relationship between a network's sparsity levels and test accuracy across different architectures with various training schedules. Overall, the results indicate little sensitivity to the training schedules, except for a slight increase in accuracy when more preliminary training is done. This increase is more pronounced for Vgg16 with TinyImageNet. For the baseline RegNetX model (i.e., 0% sparsity), the networks with only one pretraining iteration have near-zero accuracies. One pretraining iteration is not sufficient for the baseline RegNetX model. Interestingly, the accuracy for the same network increases to match the other RegNetX networks after the first iteration of pruning.

The test accuracy remains stable across all networks until 96% sparsity. This suggests that the vast majority of initial parameters in each network are redundant or non-essential to retaining accurate predictions. However, at sparsity levels exceeding 96%, the remaining parameters are essential for representing the data effectively, and further pruning leads to underparameterization. The analysis in Section IV was restricted to sparsity levels below this point, which is called the critical sparsity threshold.

Given the similarity of the accuracy vs. sparsity levels results across various retraining and training configurations, this study focuses the pruned network structure exploration on those networks with 5 preliminary training epochs and 150 training epochs. A detailed analysis of the captured parameters to explore the remaining structure of those networks is detailed in Section IV.

IV. Exploring the Pruned Structure

This section presents four experiments, summarized in Fig. 2, which explore the sparsity patterns that result from IMP. The results provide insights into why IMP is effective and the extent to which structured sparsity emerges, which has implications for memory and energy savings. To this end, IMP is applied across different architectures, as shown in Fig. 3 to prune the "prunable layers", that is, layers with trainable parameter weights such as convolutional (Conv) layers and fully connected layers (FC). Rather than using a mask, the model's weights are directly updated.

This showed that groups of contiguous layers yield nearly the same results for each experiment. Thus, to simplify the discussion and presentation of the results, similar layers are grouped into stages. Each stage consists of multiple convolutional blocks, where each block contains several layers for feature extraction followed by downsampling. These stages represent higher-level groupings of blocks that share similar characteristics (e.g., kernel size, channels, or stride). Fig. 3 shows these layer groupings. For example, Vgg16 stages consist of simple stacks of convolutional layers, ResNet stages are organized with residual blocks, and RegNetX stages use a flexible design with dynamic channel widths.

In the remainder of this section, this study will refer to stages rather than layers. Stages are referred to as either earlier or later, referring to where they occur in the CNN processing pipeline. For example, Stage 1 occurs earlier than Stage 3, as illustrated in Fig. 3.

Section IV-A explores the distribution of weights in the network being pruned, how that distribution affects pruning, and the resulting structure of the network to sparsity levels at each stage (layer) of the network relative to the overall sparsity. Section IV-B analyzes neuron sparsity. We evaluate the significance of neurons at all sparsity levels in Section IV-C. Finally, in Section IV-D we explore cosine similarity between activations of the same network across consecutive sparsity levels and the original, unpruned model.

A. Weight Distributions and Stage Sparsity

The distribution of weights across stages in the network has a significant impact on how IMP prunes the network [Fig. 2(A)]. For example, if all weights in one stage are smaller in magnitude than any weights in another stage, IMP will effectively remove the entire layer. If this occurs during the pruning process, there may be opportunities to remove those layers structurally. This section explores these distributions and how they result in later stages being pruned more than early stages in the network. As noted previously, the stage-level results are consistent with the results for individual layers within said stage.

The probability density functions (PDFs) of these weight distributions can be found in Fig. 5 (for ResNet20 and RegNetX) and Fig. 6 (for Vgg16CIFAR10 and Vgg16TinyImageNet). To simplify the presentation, four representative sparsity levels were selected: 20%, 59%, 79%, and 89%. The PDFs are scaled such that the integrals of a curve indicate the total number of weights in that stage. The later stages (higher indices) have smaller weight magnitudes but larger integrals, which implies they have many small magnitude weights. Earlier stages have fewer and larger magnitude weights.

The PDFs in Fig. 5 and Fig. 6 show that IMP initially prunes from the later stages that contain smaller magnitude

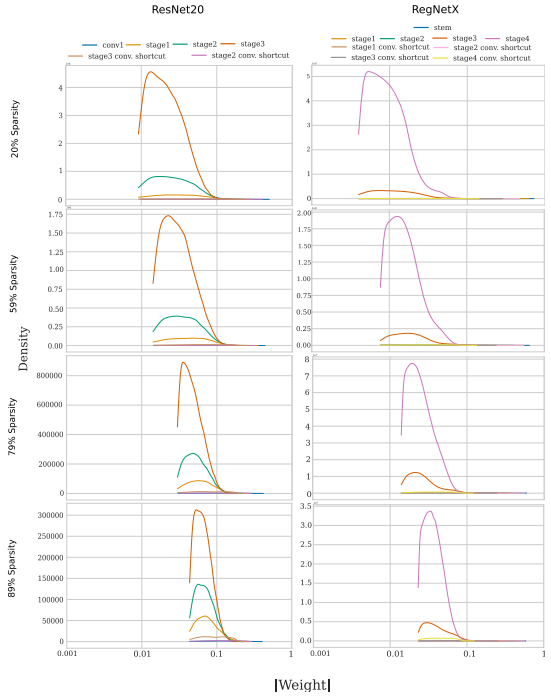


Fig. 5. The absolute value of the weight magnitude distribution across stages for varying sparsity levels for ResNet20 (Left) and RegNetX (Right).

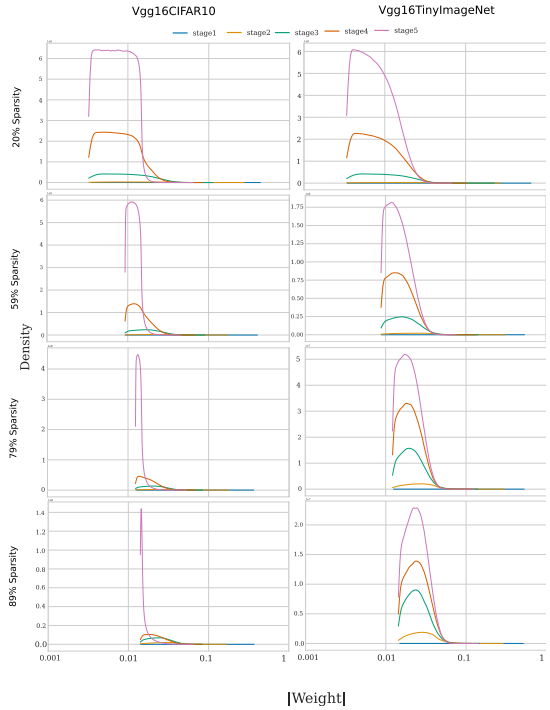


Fig. 6. The absolute value of the weight magnitude distribution across stages for varying sparsity levels for Vgg16CIFAR10 (Left) and Vgg16TinyImageNet (Right).

weights. While weights in earlier stages are still pruned, they experience less aggressive pruning, as illustrated in Fig. 7. Stage sparsity is defined as the percentage of weights pruned within a particular stage, and network sparsity as the overall percentage of pruned weights across the entire network.

We define stage sparsity as the percentage of weights in a stage that have been pruned. In contrast, the network sparsity is the percentage of weights in the *entire* network that have been pruned. In Fig. 7, the stage sparsity for each stage is plotted versus the total network sparsity. Data above or below the unit diagonal indicates that the sparsity of that stage is above or below the network average, respectively. At all network sparsities, the later stages have higher percentages of their weights pruned compared to earlier stages. This is consistent with the observations of the PDFs plotted in Fig. 5 and Fig. 6 that later stages have smaller values relative to earlier stages, resulting in later stages being more aggressively pruned. The figure illustrates these progressive sparsity dynamics up to the critical threshold, prior to the onset of accuracy degradation.

Although not plotted here, the differences in stage sparsity grow with rising network sparsity until the critical network sparsity, about 96% as mentioned previously. At that point, the sparsity of earlier stages increases rapidly. This is consistent with the degrading accuracy observed at this point in Fig. 4.

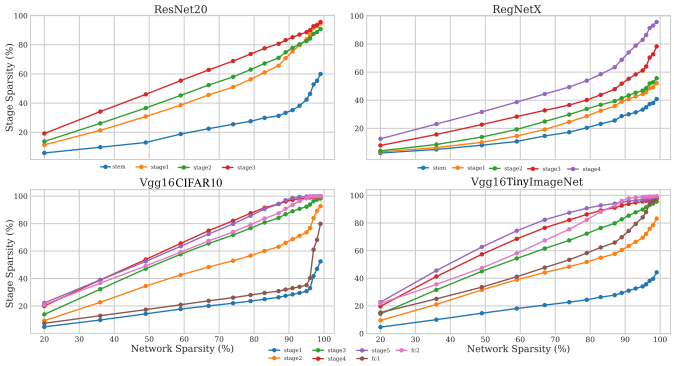


Fig. 7. **Stage-wise sparsity patterns.** Fraction of non-zero weights retained per stage and Fully Connected (FC) layers after the final sparsity level. Early stages are consistently more preserved across networks, while later stages are pruned more aggressively.

B. Neuron Sparsity

Building on insights into stage-wise pruning patterns, this study investigates the sparsity characteristics at the neuron level within these networks [Fig. 2(B)]. Given a highly sparse CNN, it is reasonable to expect that several neurons will exhibit a neuron sparsity of 100%, meaning all weights within the neuron are zeroed out. However, the experiments revealed that this was never observed. There are several reasons why this occurs. Given a constant-width layer, the probability that all the weights on an individual neuron are below the 99th percentile is given by (0.99^x), where x is the layer width. The probability that every neuron in the layer keeps at least one neuron is $(1 - 0.99^x)^x$, which tends towards 1 in an infinitely wide layer, meaning that it should not be expected to see any full neurons removed if pruning at initialization. Furthermore, it is reasonable to assume that after some weights are pruned the remaining weights in a neuron would undergo large gradient updates, further decreasing the likelihood of their removal after training. This is confirmed in Fig. 5 and Fig. 6.

This observation led to an exploration of neuron sparsity patterns within the network throughout the pruning process. The individual neuron sparsities were then grouped by layer to create layer-level neuron sparsity distributions. Tracking these distributions at each sparsity level attempts to uncover patterns in how neuron connectivity evolves. The procedure for doing this is outlined in Algorithm 1.

Algorithm 1 Layer-wise Neuron Sparsity Analysis

Require: Pruned model for all sparsity levels $\{M_s\}_{s=1}^{S}$, list of prunable layers $\mathcal{L}$

for each pruning level $s \in \{1, \ldots, S\}$ **do**
  Load network weights from checkpoint M_s
  for each layer $l \in \mathcal{L}$ **do**
    Extract weight tensor W_l of shape $[n_{\text{out}}, n_{\text{in}}]$
    Compute total weight count $T_l = n_{\text{out}} \times n_{\text{in}}$
    Compute zero weight count $Z_l = \text{count}(W_l == 0)$
    Compute layer sparsity $S_l = Z_l / T_l$
    for each output neuron $i \in \{1, \ldots, n_{\text{out}}\}$ **do**
      Compute neuron-level sparsity $s_{l,i} = \text{count}(W_l[i, :] == 0)/n_{\text{in}}$
    end for
    Store $\{S_l, \{s_{l,i}\}_{i=1}^{n_{\text{out}}}\}$
  end for
end for

Ensure: Per-layer and per-neuron sparsity statistics for all pruning sparsity levels

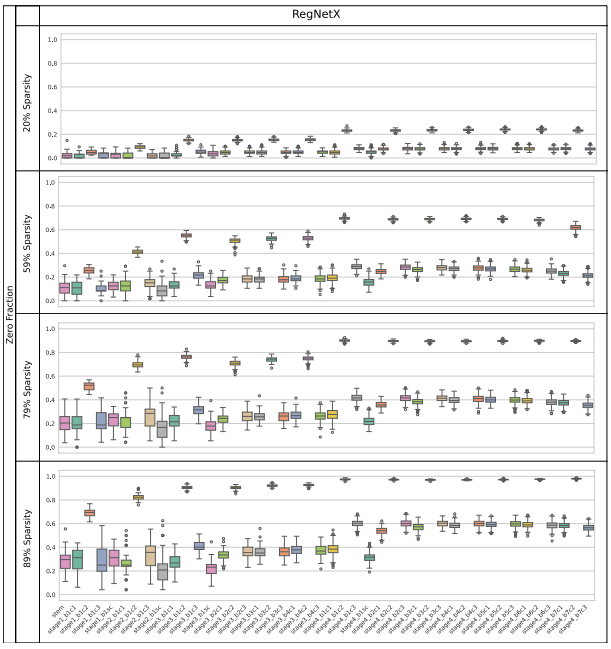


Fig. 8. Neuron sparsity distribution for each layer in RegNetX across different sparsity levels. Later layers show increased sparsity and clustering of inactive neurons, highlighting emergent structural sparsity.

Fig. 8, Fig. 9, and Fig. 10 present box-and-whisker plots of neuron sparsity across layers for each network at multiple sparsity levels. When the median neuron sparsity in a layer differs from the global network sparsity, it indicates that most neurons in that layer have been pruned more or less aggressively than average. This implies that these median values tend to increase with network depth, suggesting that neurons in later layers retain a smaller proportion of their incoming connections. This trend aligns with the stage-wise sparsity patterns shown in Fig. 7, where later layers consistently undergo heavier pruning.

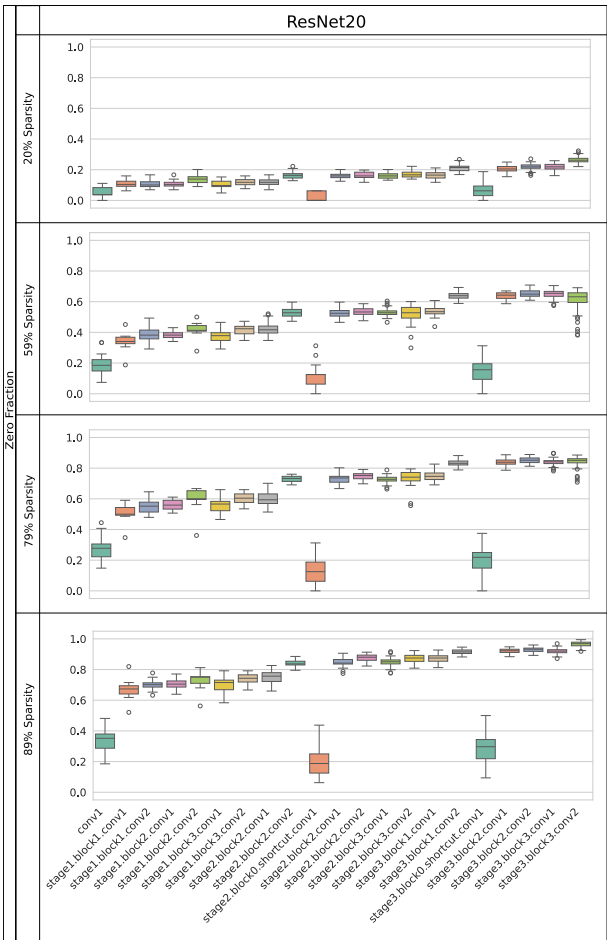


Fig. 9. Neuron sparsity distribution for each layer in ResNet20 across different sparsity levels. Later layers show increased sparsity and clustering of inactive neurons, highlighting emergent structural sparsity.

Layers that belong to the same stage, as indicated by their layer names, typically have similar neuron sparsity distributions. A notable exception to this pattern is the second convolutional layers in RegNetX, which have relatively high sparsities indicating that IMP preferentially prunes these layers. The vertical extent of the plots indicates the width of the distributions. The vertical extents are generally small compared to the variation between stages, especially for layers with high neuron sparsities. This suggests that, generally, in heavily pruned layers, all neurons are similarly sparse. However, even for the latest layers, at the highest network sparsities, no neurons exhibit exactly zero neuron sparsity.

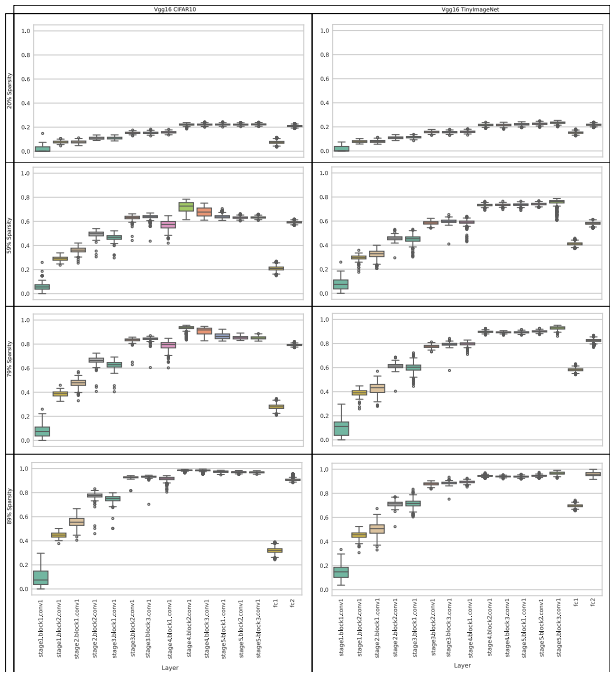


Fig. 10. Neuron sparsity distribution for each layer in Vgg16CIFAR10 and Vgg16TinyImageNet, respectively. Later layers exhibit increased sparsity and grouping of inactive neurons, indicating the presence of emergent structural sparsity.

Our neuron zeroing experiments (discussed in Section IV-C) show that removing neurons from these later layers typically results in a substantial decrease in prediction accuracy. This suggests that naive structured pruning, which removes neurons with the highest sparsity from the pruned network, would not be effective. However, this conclusion overlooks the fact that the neuron zeroing experiments do not involve retraining the networks after zeroing. Retraining following neuron removal could significantly reduce accuracy losses, potentially providing a pathway for structured pruning after IMP, but such an approach is prohibitively expensive for this investigation.

C. Neuron Zeroing

To further understand the contribution of individual neurons, a sensitivity analysis that measures the impact of neuron zeroing on prediction accuracy is conducted [Fig. 2(C)]. Neuron zeroing refers to the process of setting all weights for a specific neuron to zero. Algorithm 2 was applied to each sparsity level, with the results shown in Fig. 11 and Fig. 12.

In this work, accuracy and loss provided similar insights; since accuracy is more commonly plotted, only accuracy is reported.

Fig. 11 and Fig. 12 use box and whisker plots to show the change in accuracy distributions across all neurons. Although only the results for the first layer, two intermediate layers, and the final layer of each network are shown, all other layers show similar results.

These results show that neuron zeroing has a negligible impact on accuracy, with the highest variation being at 1.1% increase in accuracy for the unpruned Vgg16CIFAR10. For the other three cases, the sensitivity to neuron zeroing increases initially with pruning, which may indicate that as unnecessary weights are removed, each neuron is more likely to contribute favorably to the network's accuracy. This is in line with Fig. 5 and Fig. 6, where the weight magnitudes increase with each sparsity level. Then, with further pruning, rather than continuing to increase, the sensitivities remain fairly stable up to the critical sparsity.

Algorithm 2 Neuron Zeroing Experiment

Require: Trained network M, test Dataset TD, prunable layer list $\mathcal{L}$, baseline accuracy $(\text{Acc}_{\text{base}})$

```
for each layer l ∈ L do
    for each neuron i in layer l do
        Save the original weights and biases of neuron i
        Zero out weights and biases of neuron i
        Evaluate the network on TD to obtain Acc_i
        Compute ΔAcc_i = Acc_i − Acc_base
        Restore the original parameters of neuron i
    end for
end for
```

Ensure: Per-neuron sensitivity metrics for all layers

The observation that Vgg16CIFAR10 does not follow this trend may be related to the network's substantial overparameterization for the CIFAR10 dataset. This network contains 33.6 million parameters applied to a relatively low resolution dataset (32×32 image resolution). From the experimental setup, as described in Section III, the network struggles with generalization, as seen in the lower test accuracy of the unpruned network. Generalization refers to the model's ability to extract patterns from the training data and apply these patterns to unseen data, rather than memorizing specific examples. After the initial pruning step, the network shows very low sensitivity to the removal of neurons, suggesting a high degree of neuron redundancy. This behavior persists until 97% sparsity level, at which point, the model exhibits a sharp increase in neuron sensitivity. This occurs when the pruned network retains approximately 1 million parameters, which is comparable to the total parameter count of the unpruned ResNet20 (0.27 million parameters), which shows higher sensitivity from the outset of pruning. The trend is consistent with the hypothesis that, in the experiments, moderate pruning acts as a form of regularization by reducing excess capacity, thereby improving test performance before extreme pruning levels began to degrade accuracy (Fig. 4).

While retraining the network after a neuron removal could reduce accuracy loss, such an approach is computationally intractable for this scale of investigation. Furthermore, the one-shot zeroing experiments isolates the impact of each parameter, providing a strict lower-bound measure of the neuron's contribution to the network's overall accuracy.

D. Activation Similarity Across Sparsity Levels

To understand how a network's activations evolve across different levels of network sparsity using IMP, activation similarity, a measure of layer-wise activation on the same data, was analyzed across sparsity levels [Fig. 2(D)]. Specifically,

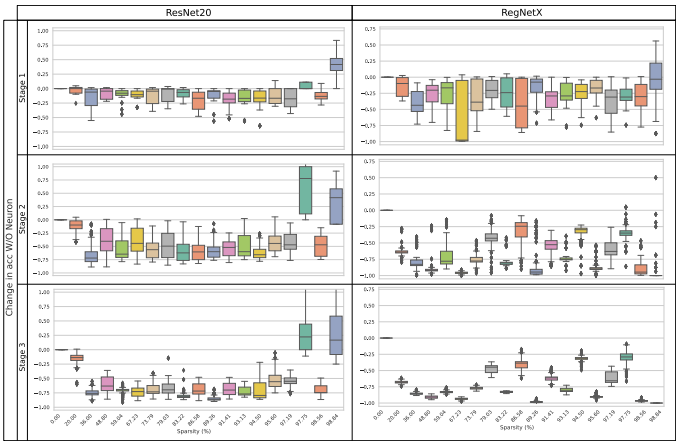


Fig. 11. The average accuracy change per stage (y-axis) from ablating individual neurons in (a) ResNet20 and (b) RegNetX networks, plotted as a function of sparsity levels (x-axis). These plots show the distribution of accuracy changes. As pruning increases, the variance and magnitude of neuron influence grow, indicating increasing functional specialization and reduced redundancy.

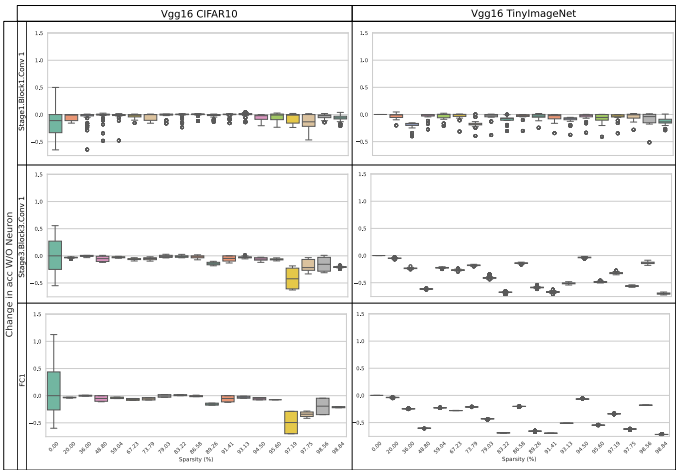


Fig. 12. Neuron zeroing impact across sparsity levels. Accuracy change (y-axis) from ablating individual neurons in (a) Vgg16CIFAR10 and (b) Vgg16TinyImageNet networks, plotted as a function of sparsity levels (x-axis). These plots show the distribution of accuracy changes. As pruning increases, the variance and magnitude of neuron influence grow, indicating increasing functional specialization and reduced redundancy.

the activation outputs of every neuron are recorded on the test set during inference at each sparsity level (after retraining the pruned network). Pairwise activation similarity matrices are then computed for the same layer across different sparsity levels using cosine similarity [Eq. (1)] to quantify changes in representation.

To compute the similarity matrices for each layer, neuron outputs are reshaped into two-dimensional matrices of shape $[\text{samples}, \text{neurons}]$. In convolutional layers, where each "neuron" corresponds to a filter, activations are first flattened to $[\text{samples} \times \text{width} \times \text{height}, \text{filters}]$ before computing cosine similarity. Algorithm 3 details this process.

The computed cosine similarity of layer-wise activation from a given sparsity level to the baseline network is shown

Algorithm 3 Activation Similarity Calculation

Require: Trained network M, test Dataset TD, prunable layer list $\mathcal{L}$, sparsity level s

1: Load network weights at sparsity level s
2: Run inference of M_s on TD, collecting activations A_l for each layer
3: Run inference of M_{s-1} or M_0 on TD, collecting activations B_l for each layer.
4: **for** each layer l **do**
5: Reshape A_l and B_l to $[\text{samples}, \text{neurons}]$
6: Compute similarity score S_l using Eq. (1).
7: **end for**

Ensure: $\{S_l\}_{l \in \mathcal{L}}$

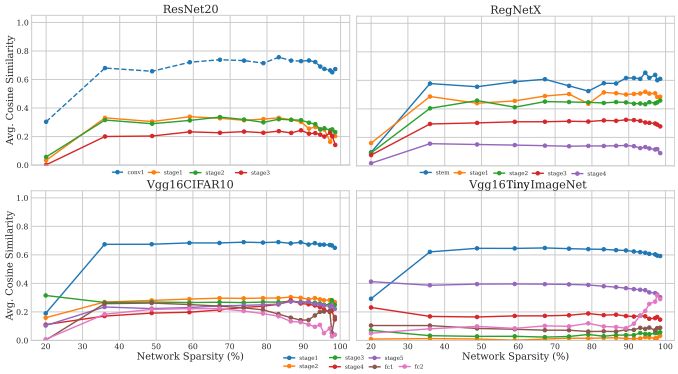


Fig. 13. Activation similarity to baseline network. The cosine similarity between layer-wise activations in the pruned network and the baseline network. This figure shows that early layers retain high similarity even at high sparsity levels, while later layers share little similarity.

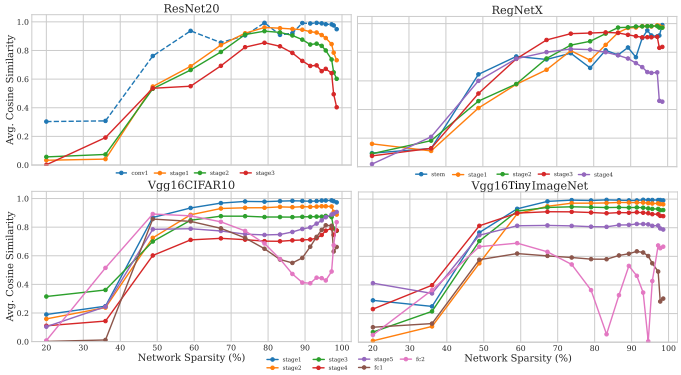


Fig. 14. Step-wise activation similarity across pruning iterations. The cosine similarity between activations of the same network across sparsity levels. The similarity declines gradually with pruning, but drops sharply near the critical sparsity level, signaling large degeneration in the network's representation.

in Fig. 13. It is observed that activation similarities between the sparsity level and the baseline decrease as the stage index increases. That is, later stages are less similar to the baseline than early stages. This is likely because the activations of later stages depend on those of earlier stages, thereby compounding the dissimilarity in subsequent stages.

This persistent similarity of activations to their values in the baseline network may imply that each pruned network consistently finds a similar basin of attraction in the loss landscape. In fact, pruning may increase the likelihood of finding the same basin of attraction by further constraining the solution as weights are set to zero.

The main exception to this activation similarity observation is at a network sparsity of 20%, where the similarities are substantially lower than at other points. This drop in similarity is observed (but not included in the manuscript) even with various pretraining and training epochs. As shown in Fig. 4, the accuracies at 20% network sparsity with 150 training epochs are not particularly low, so this suggests an alternate local minimum in the loss landscape has been identified. More investigation would be required to determine why such alternate minima are not identified at other network sparsities.

Next, similarities are considered between activations of the same network across sparsity levels, as shown in Fig. 14. The first two pruning iterations (20% and 36% sparsity) show that the activation function outputs change drastically between iterations. This is consistent with Fig. 13, where the first pruning iteration has low similarity to the baseline and the second has high similarity. Thus, the 20% sparsity level network is stage-wise dissimilar to the baseline, and then the next iteration is similarly dissimilar to the first iteration (and more similar to the baseline).

Subsequent stage-wise similarities are much higher, as they are similar to the second iteration (36% sparsity). Many of the similarities are even above 60% or even 80%, which is remarkable considering the high dimensionality of the vectors considered. This indicates that the neuron activations become highly similar between pruning iterations for all layers, with the highest similarities observed for the earliest convolutional layers. The stage-wise similarities of the convolutional layers remain high until the critical network sparsity. For the fully connected layers in Vgg, the stage-wise similarities begin to decrease well before the critical sparsity. This may be because these layers are the latest in the architecture and must adapt to changes in earlier layers.

V. Conclusion

This study explores the structure resulting from using IMP to prune a series of CNNs. A detailed analysis of IMP of both neurons and layers across three CNN architectures (ResNet20, RegNetX, and Vgg16) and two datasets (CIFAR10 and TinyImageNet) was conducted. We arrive at four key observations:

- Preferential pruning of deep layers Because CNNs concentrate more parameters in deeper layers and weight magnitudes typically scale inversely with layer width, IMP naturally removes more weights from late stages. This preserves early convolutions, critical for capturing fundamental image features, and explains why accuracy remains high up to extreme sparsities, only degrading as early layers are pruned substantially. The preferential pruning of deep layers suggests that inference performance may be improved by using differing data structures in various parts of the network (e.g., dense matrices in early stages and sparse matrices in later stages).
- Persistence of partially sparse neurons While many neurons retain high sparsity late into pruning, none reach full zeroing. The effect of these "last few weights" limits potential matrix-size reduction and, hence, hardware efficiency. The findings motivate IMP variants that explicitly target the removal of residual weights within neurons.
- Stable neuron-level sensitivity distributions Contrary to expectations, the variance of single-neuron importance does not narrow as pruning progresses: remaining neurons continue to exhibit a broad range of sensitivities. It is observed that one-shot zeroing experiments (without retraining) may overestimate individual neuron importance; future work could apply iterative zeroing with fine-tuning to refine these insights.
- Preservation of feature representations Cosine-similarity analysis show that IMP minimally perturbs activation patterns throughout pruning. This suggests pruned and unpruned networks occupy nearby regions of parameter space, maintaining the representational capacity that underlies accuracy preservation.

Overall, these observations clarify why IMP achieves very high unstructured sparsities without sacrificing accuracy, and they highlight why magnitude-based pruning alone is unlikely to yield structured sparsity patterns conducive to hardware acceleration. Future experiments hope to: 1) develop IMP extensions that incentivize fully zeroing neurons, 2) evaluate the runtime and energy impacts on real hardware, and 3) extend the analysis to other model families (e.g., transformers and graph networks).

Conflicts of Interest

The authors have no conflicts of interest to declare.

Data Availability Statement

The software developed for this research is available at the following link:

https://github.com/NatLabRockies/network-pruner

The data presented in this manuscript can be made available upon reasonable request.

Acknowledgment

This work was authored in part by the National Laboratory of the Rockies (NLR) for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported in part by the Laboratory Directed Research and Development (LDRD) Program at NLR. The views expressed

in the study do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the study for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. A portion of the research was performed using computational resources sponsored by the Department of Energy's Office of Critical Minerals and Energy Innovation and located at the National Laboratory of the Rockies.

REFERENCES

[1] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," *Advances in neural information processing systems*, vol. 2, 1989.

[2] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.

[3] M. C. Mozer and P. Smolensky, "Skeletonization: A technique for trimming the fat from a network via relevance assessment," *Advances in neural information processing systems*, vol. 1, 1988.

[4] Sietsma and Dow, "Neural net pruning-why and how," in *IEEE 1988 international conference on neural networks*. IEEE, 1988, pp. 325–333.

[5] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. De Freitas, "Predicting parameters in deep learning," *Advances in neural information processing systems*, vol. 26, 2013.

[6] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: efficient inference engine on compressed deep neural network," in *Proceedings of the 43rd International Symposium on Computer Architecture*, ser. ISCA '16. IEEE Press, 2016, p. 243–254. [Online]. Available: https://doi.org/10.1109/ISCA.2016.30

[7] Z. Yang and H. Zhang, "Comparative analysis of structured pruning and unstructured pruning," in *International Conference on Frontier Computing*. Springer, 2021, pp. 882–889.

[8] A. Bragagnolo, E. Tartaglione, A. Fiandrotti, and M. Grangetto, "On the role of structured pruning for neural network compression," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3527–3531.

[9] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.

[10] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," *arXiv preprint arXiv:1902.09574*, 2019.

[11] J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin, "Linear mode connectivity and the lottery ticket hypothesis," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 3259–3269. [Online]. Available: https://proceedings.mlr.press/v119/frankle20a.html

[12] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient transfer learning. arxiv 2016," *arXiv preprint arXiv:1611.06440*, 2016.

[13] J. Pieterse and D. Mocanu, "Evolving and understanding sparse deep neural networks using cosine similarity," *ArXIV*, vol. 2019, pp. 1903–07 138, 2019.

[14] A. RoyChowdhury, P. Sharma, E. Learned-Miller, and A. Roy, "Reducing duplicate filters in deep neural networks," in *NIPS workshop on deep learning: Bridging theory and practice*, vol. 1, no. 2, 2017, p. 6.

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

[16] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 428–10 436.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[19] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[20] F. O. Giuste and J. C. Vizcarra, "Cifar-10 image classification using feature ensembles," 2020. [Online]. Available: https://arxiv.org/abs/2002.03846

[21] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," 2020. [Online]. Available: https://arxiv.org/abs/2003.13678

Privacy Leakage and Memorization in Fine-Tuned Clinical Language Models: A Controlled Study of Defenses and Backbone Choice on Clinical Narrative Transcriptions

Yassine Chahid[1], Anas Chahid[2], Ismail Chahid[3], Aissa Kerkour Elmiad[4]
ACSA, Mohammed First University, Oujda, 60000, Morocco[1]
SmartICT, Mohammed First University, Oujda, 60000, Morocco[2]
LARI, Mohammed First University, Oujda, 60000, Morocco[3,4]

***Abstract*—The increasing adoption of large language models (LLMs) and domain-adapted transformers in healthcare has created a new privacy challenge: fine-tuned models may memorize rare clinical strings and later reveal them through generation or scoring behavior. A controlled study of privacy leakage and memorization in clinical language models trained on narrative transcriptions is presented. A canary-based audit pipeline was instantiated on a 4,000-note subset of the Medical Transcriptions (MTSamples) corpus, with 40 synthetic secrets injected only into the training partition and evaluated using three complementary attack families: prompt extraction, exposure-style ranking, and reference-based membership inference. Two experiments are reported. Experiment I compares baseline fine-tuning, early stopping, and a conservative regularized training profile combining lower learning rate, higher weight decay, and partial layer freezing. Experiment II fixes the training protocol and compares DistilGPT2, GPT-2, and BioGPT. A clear privacy-utility tension was observed. In Experiment I, early stopping produced the best held-out language-model utility, whereas the combined regularized profile eliminated observed prompt leakage and reduced membership-inference strength, at the cost of worse perplexity. In Experiment II, stronger and more domain-specialized backbones achieved better clinical language modeling but also exhibited higher leakage and stronger membership-inference signals, with BioGPT yielding the strongest utility and the highest privacy risk under the evaluated attacks. These results indicate that privacy auditing should accompany utility evaluation in clinical LLM adaptation, and that backbone choice can materially affect memorization risk in this controlled setting.**

***Keywords*—Clinical language models; privacy leakage; memorization; membership inference; canary exposure; BioGPT; GPT-2; healthcare NLP; model auditing**

I. Introduction

Large language models are rapidly becoming part of healthcare AI pipelines, including clinical decision support, summarization of electronic health records (EHRs), documentation assistance, patient communication, and biomedical question answering [2], [3], [4], [5], [6]. Their appeal is clear: once adapted to clinical language, such models can absorb domain vocabulary, structure, and reasoning patterns that are costly to encode manually. At the same time, their deployment introduces an acute privacy question: *when a general or biomedical backbone is fine-tuned on clinical text, does it merely generalize, or does it also memorize rare sensitive sequences strongly enough to leak them later?*

This question is especially pressing in healthcare because patient narratives often contain sparse, low-frequency identifiers and clinically meaningful attributes. Even when direct identifiers are removed, combinations of age, disease code, specialty, procedural detail, and treatment narrative can become quasi-identifying. Responsible machine learning in medicine therefore requires more than utility evaluation alone; it requires explicit privacy auditing and leakage measurement [1], [7], [33], [8].

Privacy concerns in language models are not hypothetical. Work on unintended memorization, training-data extraction, and membership inference has shown that generative models can reveal exact or near-exact traces of the data seen during training [9], [10], [11], [12], [13], [14]. Recent findings further suggest that stronger or more domain-adapted models may become more data efficient but simultaneously more privacy sensitive [15], [16], [17], [19].

Clinical adaptation adds another layer of complexity. Biomedical and clinical language models such as BioBERT, PubMedBERT, GatorTron, Med-BERT, BEHRT, and BioGPT outperform general models on many healthcare tasks [26], [27], [29], [30], [31], [28]. However, stronger in-domain fit may also increase the probability of memorizing rare strings. The literature has discussed this possibility conceptually, but practical, reproducible demonstrations on clinical-style corpora are still limited, especially when both defenses and backbone choice are evaluated within the same study.

That gap is addressed here through a controlled dual-experiment study on clinical narrative transcriptions. The Medical Transcriptions (MTSamples) dataset is used as a multi-specialty corpus for clinical free text, and a canary-based audit protocol is designed in which synthetic secrets are inserted *only into the training split*. The resulting fine-tuned models are then tested for leakage through generation and for memorization through lower loss and higher ranking scores (Fig. 1).

Four contributions are made:

- A defense-oriented privacy audit. Baseline fine-tuning, early stopping, and a combined regularized training

profile are compared on the same controlled clinical fine-tuning task.

- A backbone-oriented privacy audit. DistilGPT2, GPT-2, and BioGPT are compared under the same dataset, split, canary set, and attack suite.
- A multi-attack evaluation protocol. Leakage is quantified with prompt-based extraction, candidate ranking with exposure-style scoring, and reference-based membership inference using ROC-AUC, PR-AUC, accuracy, precision, recall, F1, specificity, and balanced accuracy.
- A defensible privacy and utility analysis. The reported results indicate that stronger utility does not guarantee better privacy, and that backbone choice is itself a privacy-relevant design decision in clinical LLM fine-tuning.

II. Related Work

A. Healthcare Language Models and Clinical Adaptation

Clinical and biomedical transformer models have progressed rapidly, from early domain adaptation strategies to large-scale specialized pretraining. BioBERT improved biomedical text mining with domain-adapted contextual representations [26]. PubMedBERT showed that in-domain pretraining from scratch can outperform general-domain initialization in biomedical NLP [27]. Structured-clinical representation learning advanced through Med-BERT and BEHRT, which demonstrated that transformer-based pretraining on large EHR repositories benefits downstream prediction tasks [30], [31]. At larger scale, GatorTron and BioGPT extended this line to generative and clinical-free-text modeling [29], [28]. More recently, Med-PaLM-style systems and health system-scale language models have highlighted the emerging power of foundation models in healthcare [4], [5], [6], [41].

B. Privacy Risks in Machine Learning and Clinical AI

The healthcare literature has repeatedly emphasized that high utility alone is insufficient in medical AI. Responsible deployment requires privacy, governance, reproducibility, and harm-aware design [1], [2], [3]. In medical imaging and distributed healthcare learning, privacy-preserving federated methods and secure training protocols have become major themes [7], [32], [33], [34], [35], [8]. Yet privacy issues in generative clinical text models remain comparatively underexplored, especially under white-box or score-based audit settings.

C. Memorization, Training-Data Extraction, and Membership Inference

Membership inference emerged as a canonical privacy threat in machine learning through black-box attacks that infer whether a sample belonged to the training set [11]. In language models, memorization and extraction risks became especially visible with canary-based audits and direct generation attacks [9], [10]. Later work showed that deduplication, training dynamics, and neighborhood-based attacks strongly affect leakage behavior [15], [16], [12], [13], [17], [18], [19]. Recent work has extended the threat model to code models, context-aware membership inference, unlearning of personally identifiable information, and broad privacy-risk surveys for LLMs [20], [21], [22], [25], [23], [24].

D. Differential Privacy and Related Defenses

Differential privacy remains the most principled formal defense against memorization, beginning with DP-SGD [36]. Its use in language modeling and large-model fine-tuning is active but computationally expensive, and often introduces a non-trivial utility penalty [37], [38], [39]. Outside formal privacy, practical mitigations include data deduplication, regularization, early stopping, dropout, freezing submodules, prompt filtering, and unlearning [15], [16], [22]. However, these defenses are rarely compared side by side on clinical-style generative fine-tuning under the same controlled attack suite.

The study is positioned at this intersection: clinical-language fine-tuning, memorization-aware auditing, defense comparison, and backbone comparison, all grounded on the same clinical narrative corpus.

III. Materials and Methods

A. Dataset and Sampling Protocol

Both experiments were conducted on the Medical Transcriptions (MTSamples) dataset, available at https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions. The corpus contains 4,999 clinical-style transcriptions spanning 40 specialty categories. After preprocessing, 4,000 notes were sampled. The resulting splits followed an 80/10/10 partition: 3,200 training notes, 400 validation notes, and 400 test notes. For both experiments, exact duplicates were removed before splitting.

The selected corpus is appropriate for privacy auditing because it contains long-form narrative prose, specialty-specific vocabulary, structured reporting habits, and heterogeneous note styles. These properties are important because memorization risk is driven not only by raw dataset size, but also by the presence of rare phrasings, repeated templates, and clinically meaningful identifiers embedded in free text.

TABLE I. Dataset and Core Protocol Settings Used in Both Experiments

Setting	Value
Base corpus	Medical Transcriptions (MTSamples)
Prepared note count	4,000 sampled notes
Train/validation/test split	3,200 / 400 / 400
Sequence length	128 tokens
Batch size	2
Gradient accumulation	4
Injected canaries	40
Control canaries	40
Canary repetitions in training	2
Prompt families	4
Candidate pool for ranking attack	64 condition codes

B. Synthetic Canary Design

To measure memorization under controlled conditions, synthetic canaries of the form below were injected:

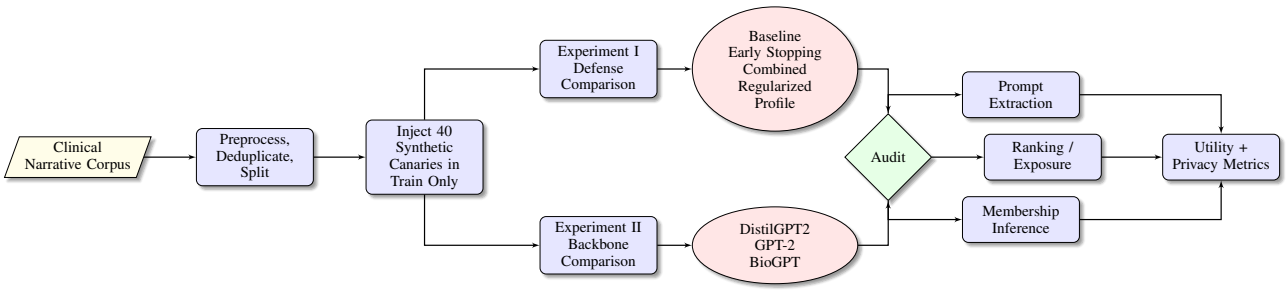


Fig. 1. Overall study design. A clinical narrative corpus is sampled, preprocessed, split, and augmented with train-only synthetic canaries. Experiment I compares defenses on one backbone, while Experiment II compares backbones under a fixed training protocol. All models are audited with prompt extraction, ranking and exposure analysis, and reference-based membership inference.

```
Patient identifier ZXQ-XXXX has condition
OMEGA-YY.
```

Each patient identifier and condition code was unique. The canaries were embedded in realistic clinical-note fragments and inserted *only into the training split*. An additional set of 40 control canaries was generated with the same syntactic structure but never injected into the model. This control set enabled membership inference and ranking-based discrimination between members and non-members.

C. Experiment I: Defense Comparison

Experiment I used DistilGPT2 as the backbone and compared three training settings:

- Baseline: 3 epochs, learning rate 5×10^{-5}, weight decay 0.01.
- Early stopping: up to 6 epochs with patience 2, same optimizer scale as baseline.
- Combined regularized profile: 4 epochs maximum, learning rate 2×10^{-5}, weight decay 0.05, warmup ratio 0.05, and freezing of the lower half of the transformer stack.

Because this third setting combines several interventions, the resulting privacy and utility changes should be interpreted at the profile level rather than attributed to any single component in isolation.

D. Experiment II: Backbone Comparison

Experiment II fixed the training protocol and compared three backbones:

- DistilGPT2 (81.9M parameters),
- GPT-2 (124.4M parameters),
- BioGPT (346.8M parameters).

All three models were trained for up to 4 epochs with learning rate 5×10^{-5}, weight decay 0.01, warmup ratio 0.03, and early stopping patience 2. This made the comparison as controlled as possible: only the backbone changed.

E. Experimental Architecture and Reproducibility Controls

Both notebooks implemented the same end-to-end experimental architecture. First, the corpus was downloaded, normalized, and partitioned into train, validation, and test splits with fixed seeds. Second, synthetic canaries were generated automatically and inserted only into the training split, while matched control canaries were retained as non-members. Third, a chosen causal-language-model backbone was fine-tuned under an explicitly declared profile. Fourth, the trained model was audited by three privacy-evaluation layers: prompt extraction, ranking-based memorization analysis, and reference-based membership inference. Finally, every run exported CSV summaries, plots, and serialized artifacts for downstream inspection and manuscript generation.

Several design choices strengthen scientific defensibility. The experiments used the same prompt templates across all runs, fixed train/validation/test boundaries, CPU-compatible batch sizes, explicit configuration blocks, persistent result files, and no post-hoc retuning of the privacy protocol after observing outcomes. In the backbone-comparison notebook, hyperparameters were held constant across backbones so that privacy differences could be interpreted with minimal optimization asymmetry. This improves comparability, although it does not eliminate configuration sensitivity. Reproducibility is further supported through fixed random seeds, persistent exported artifacts, and complete reporting of the principal training settings in Tables I, III, and IV. Software-version pinning, runtime-variance profiling across repeated executions, and memory-usage tracking were not analyzed in this study and should therefore be considered outside the present reproducibility envelope.

F. Attack-Suite Design

The attack suite was designed to cover multiple black-box interaction patterns rather than a single prompt template. Leakage can surface through chart-style lookups, incomplete clinical notes, routine continuation requests, or repeated model queries. For that reason, the notebooks tested several prompt families rather than one handcrafted query. Greedy decoding alone can underestimate risk, so beam-search and stochastic-sampling variants were also included. This mattered empirically: in Experiment I, some canaries only became exact leaks under sampled decoding, while greedy and beam decoding produced only partial continuations.

The ranking and membership attacks deepen the analysis beyond visible completions. Exposure-style ranking does not require a verbatim leak; it asks whether the true secret is scored unusually well relative to many plausible alternatives. Membership inference asks a related question: can an adversary infer whether a sensitive pattern was used during fine-tuning? In healthcare, even that signal can be consequential because it may reveal that a rare condition template, documentation style, or institution-specific phrase was present in the fine-tuning distribution. Taken together, the three attack families provide a stronger audit than prompt extraction alone.

G. Attack Suite

Three families of privacy attacks were used.

1) Prompt-based extraction: For each canary, four prompt families were evaluated:

- prefix completion,
- direct recall,
- partial note completion,
- chart lookup.

Each prompt was decoded with greedy decoding, beam search, and stochastic sampling. The prompt-based metrics were:

- exact-match recovery rate,
- condition recovery rate,
- patient-identifier recovery rate,
- partial-match rate,
- token recovery rate,
- average string similarity,
- leakage rate (fraction of canaries leaked at least once).

2) Ranking and exposure: Prompt attacks can underestimate memorization because a secret may be internally preferred without always appearing under one decoding strategy. Multiple candidate condition codes were therefore scored given a fixed patient identifier, and the true code was ranked among 64 candidates. Exposure-style memorization was computed as:

$$E = \log_2(N) - \log_2(r), \tag{1}$$

where, N is the candidate-pool size and r is the rank of the true secret. Higher exposure indicates stronger memorization.

3) Reference-based membership inference: For each fine-tuned model, a reference model with the same base weights before fine-tuning was retained. Each member and non-member canary note was scored by both the target model and the reference model. The primary membership score was the loss gap:

$$\Delta\mathcal{L} = \mathcal{L}_{\text{reference}} - \mathcal{L}_{\text{target}}. \tag{2}$$

ROC-AUC, PR-AUC, accuracy, precision, recall, F1, specificity, and balanced accuracy were then evaluated.

H. Utility Metrics

To assess language-model quality independently of privacy, validation loss, test loss, and perplexity were reported:

$$\text{Perplexity} = \exp(\mathcal{L}), \tag{3}$$

where, $\mathcal{L}$ is the mean next-token cross-entropy loss on the corresponding split.

I. Statistical Uncertainty

The model-comparison notebook additionally computed bootstrap confidence intervals for leakage rate, exposure, and ROC-AUC, as well as paired Wilcoxon signed-rank tests on per-canary exposure values. The bootstrap intervals are used descriptively to summarize uncertainty around the reported metrics, whereas the Wilcoxon tests are used only for cautious pairwise interpretation of exposure differences (Table II).

IV. Results: Experiment I: Defense Comparison

A. Training Dynamics and Utility

The three training settings converged cleanly, but their utility profiles diverged markedly. As shown in Fig. 2, early stopping achieved the strongest held-out language modeling with validation loss 1.770 and validation perplexity 5.87, outperforming the baseline (validation perplexity 7.10). The combined regularized profile substantially weakened language-model fit, yielding validation perplexity 10.46 and test perplexity 10.36.

Fig. 2 shows that the training curves do not simply indicate optimization success; they also frame the privacy interpretation. The early-stopping run descends below the baseline on validation loss and preserves stable generalization behavior, confirming that it is the strongest utility configuration in Experiment I. The combined regularized profile, by contrast, converges more conservatively and remains consistently higher on validation loss, which is consistent with stronger capacity control.

The privacy summary in the companion panel is equally important. It shows that the configuration with the best language-model fit is not the one with the best privacy behavior. This visual contrast is central to the study's argument because it demonstrates that utility metrics alone would have selected a more privacy-sensitive model.

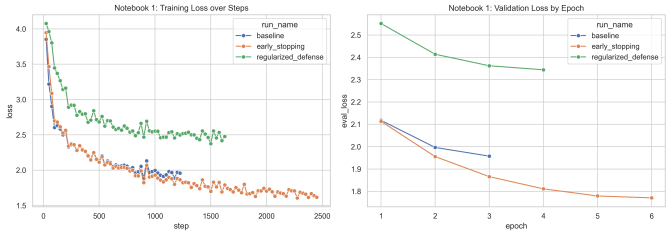
B. Prompt Leakage

The prompt-based results show that utility improvements did not translate into better privacy. Baseline and early stopping both leaked 7 canaries out of 40, corresponding to a leakage rate of 0.175. The combined regularized profile reduced the number of leaked canaries to zero. Exact-match recovery was also lowest for that profile (0.000), compared with 0.0075 for the baseline and 0.01375 for early stopping.

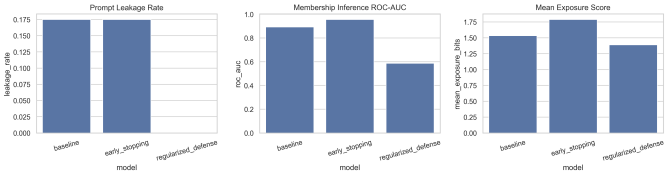
An important nuance is that the patient-identifier recovery rate saturated at 1.0 for all settings because the prompts explicitly included the patient identifier. Accordingly, the most meaningful prompt-level privacy indicators in this study are

TABLE II. ATTACK TAXONOMY AND EVALUATION METRICS USED THROUGHOUT THE STUDY

Attack family	Operational idea	Primary metrics
Prompt extraction	Generate continuations from clinically plausible prompts	Exact match, condition recovery, leakage rate
Ranking / exposure	Rank the true condition against 64 candidate codes	Mean rank, top-1 success, exposure bits
Membership inference	Distinguish seen from unseen canary notes using loss-gap scores	ROC-AUC, PR-AUC, F1, specificity, balanced accuracy



(a) Training loss and validation loss for Experiment I.



(b) Privacy summary for Experiment I.

Fig. 2. Experiment I results. Early stopping gave the best utility, whereas the combined regularized profile most strongly reduced the observed privacy risk metrics.

the condition recovery rate, exact-match recovery rate, and per-canary leakage rate. On those metrics, early stopping slightly worsened privacy relative to the baseline, while the combined regularized profile eliminated observed prompt leakage. Because the profile bundles lower learning rate, higher weight decay, and lower-layer freezing, these gains should be interpreted as the effect of the combined training configuration rather than as evidence for any single defense mechanism.

C. Ranking and Membership Inference

Ranking-based memorization and membership inference reinforce the same conclusion. Early stopping had the highest mean exposure score (1.79 bits) and the strongest membership-inference signal (ROC-AUC 0.956), whereas the combined regularized profile had the lowest mean exposure score (1.39 bits) and a much weaker membership signal (ROC-AUC 0.586). The baseline was intermediate, with ROC-AUC 0.893 and mean exposure 1.53 bits.

Experiment I therefore indicates that the strongest utility configuration was also the most privacy-sensitive, while the combined regularized profile reduced the observed leakage metrics at the cost of substantially worse clinical language modeling. In other words, in this setting, early stopping was not a privacy defense even though it remained a useful optimization strategy (Fig. 3).

The attack-type breakdown adds another layer of interpretation. Direct-recall prompts were the least effective across settings, whereas prefix completion and partial-note completion were more likely to reveal condition codes. This matters because it shows that privacy leakage is tied closely to continuation behavior in clinically shaped contexts. In practical terms, the attack surface is not limited to explicit question prompts; it extends to standard completion workflows that resemble documentation assistance or chart continuation.

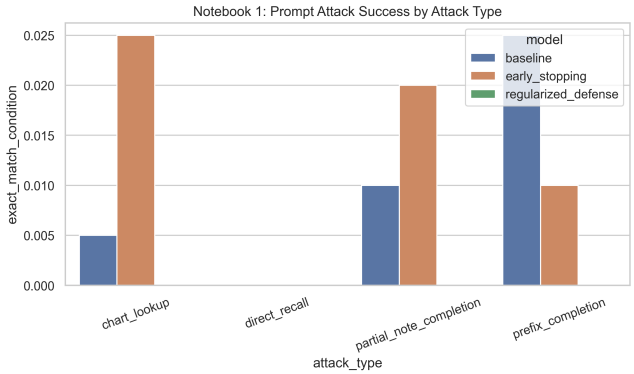


Fig. 3. Prompt attack success by attack type in Experiment I. Prefix-style and partial-note completions were more effective than direct recall prompts.

V. RESULTS: EXPERIMENT II: BACKBONE COMPARISON

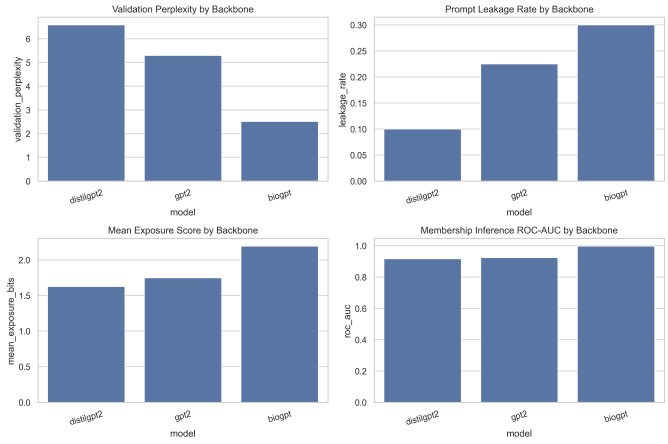
A. Utility Comparison Across Backbones

The model-comparison study yielded a clear performance ordering. DistilGPT2 achieved validation perplexity 6.59 and test perplexity 6.55. GPT-2 improved on this with validation perplexity 5.30 and test perplexity 5.33. BioGPT substantially outperformed both, reaching validation perplexity 2.52 and test perplexity 2.60. The training curves in Fig. 4 show faster and deeper optimization for BioGPT than for GPT-2 or DistilGPT2.

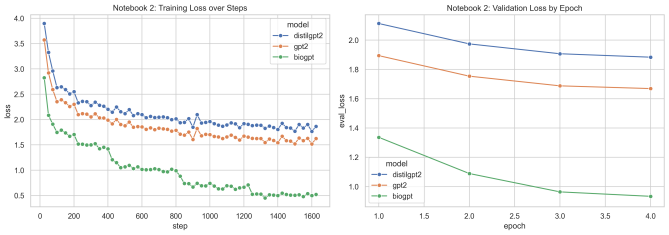
The two panels in Fig. 4 deserve joint interpretation. The utility panel shows a sharp improvement from DistilGPT2 to GPT-2 and then to BioGPT, especially in perplexity. The training-curve panel complements this by showing that BioGPT does not merely finish with a lower loss; it reaches a lower validation regime earlier and maintains that advantage across epochs. This pattern is consistent with a backbone that starts closer to the biomedical target distribution and adapts more efficiently to the corpus.

TABLE III. DEFENSE COMPARISON RESULTS (EXPERIMENT I)

Setting	Val. PPL	Test PPL	Leakage	Exact Match	Exposure	ROC-AUC	PR-AUC	F1
Baseline	7.10	7.05	0.175	0.0075	1.533	0.893	0.895	0.831
Early stopping	5.87	5.88	0.175	0.0138	1.788	0.956	0.953	0.907
Combined regularized profile	10.46	10.36	0.000	0.0000	1.388	0.586	0.537	0.692



(a) Utility and privacy summary across backbones.



(b) Training and validation curves across backbones.

Fig. 4. Experiment II results. BioGPT achieved the best utility, but also the highest leakage, exposure, and membership-inference risk.

B. Prompt Leakage and Memorization

The privacy ordering was the inverse of the "safest utility" interpretation. DistilGPT2 had the lowest leakage rate (0.10, i.e., 4 leaked canaries out of 40), GPT-2 leaked 9 canaries (0.225), and BioGPT leaked 12 canaries (0.30). Exact-match recovery rate rose from 0.00625 for DistilGPT2 to 0.02344 for GPT-2 and 0.03594 for BioGPT. Likewise, condition recovery rate increased monotonically across the same ordering.

Ranking-based exposure confirmed this pattern. Mean exposure rose from 1.63 bits for DistilGPT2 to 1.75 bits for GPT-2 and 2.19 bits for BioGPT. The average rank of the true secret improved (that is, privacy worsened) from 26.1 for DistilGPT2 to 25.1 for GPT-2 and 19.4 for BioGPT. The best-performing clinical backbone therefore also ranked the true secret highest among decoys.

C. Membership Inference

Membership inference produced the strongest privacy signal in Experiment II. DistilGPT2 and GPT-2 already had very high ROC-AUC values (0.918 and 0.925), but BioGPT approached perfect separation with ROC-AUC 0.998 and PR-AUC 0.998. The corresponding 95% bootstrap confidence interval for BioGPT was [0.991, 1.000], making it highly unlikely that the observed effect is a numerical artifact. This finding means that, under the evaluated loss-gap attack, the fine-tuned BioGPT almost perfectly distinguished seen canary notes from unseen controls.

D. Attack-Type Analysis

The backbone comparison also revealed prompt-family differences. Direct recall prompts were the weakest overall, whereas prefix completion, chart lookup, and partial-note completion were more successful. BioGPT dominated these attack families, suggesting that clinically plausible continuation contexts are especially effective at surfacing memorized strings from stronger biomedical backbones.

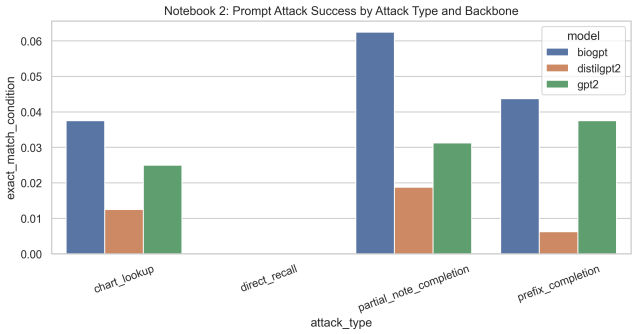


Fig. 5. Prompt attack success by attack type and backbone. Clinically natural completion prompts were more effective than direct recall prompts, especially against BioGPT.

This figure (Fig. 5) is useful not only because BioGPT is highest in aggregate, but also because the ordering is stable across attack families. That stability reduces the likelihood that the result is an artifact of one particular prompt wording. Instead, it suggests that the observed privacy ordering is not confined to a single prompt family.

E. Confidence Intervals and Paired Statistics

Bootstrap intervals preserved the same descriptive ordering. DistilGPT2 had a leakage-rate interval of [0.025, 0.200], GPT-2 had [0.075, 0.375], and BioGPT had [0.175, 0.450].

TABLE IV. BACKBONE COMPARISON RESULTS (EXPERIMENT II)

Backbone	Val. PPL	Test PPL	Leakage	Exact Match	Exposure	ROC-AUC	PR-AUC	F1	Train Time (s)	Params
DistilGPT2	6.59	6.55	0.100	0.0063	1.627	0.918	0.918	0.843	5226	81.9M
GPT-2	5.30	5.33	0.225	0.0234	1.748	0.925	0.922	0.822	7842	124.4M
BioGPT	2.52	2.60	0.300	0.0359	2.194	0.998	0.998	0.974	13588	346.8M

Membership ROC-AUC intervals also separated the models, with BioGPT concentrated near one. The paired Wilcoxon analysis on per-canary exposure values showed directionally higher exposure for BioGPT than for the other backbones, although at this sample size the pairwise p-values (0.086 to 0.100) remained above 0.05. DistilGPT2 and GPT-2 were not significantly different on exposure under the same test ($p = 0.615$).

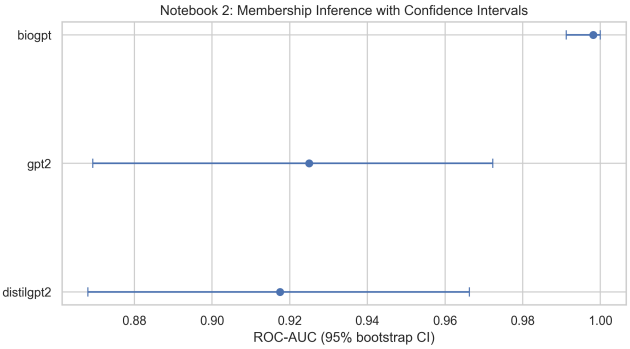


Fig. 6. Membership-inference ROC-AUC with 95% bootstrap confidence intervals in Experiment II. BioGPT exhibits near-perfect separability between member and non-member canary notes.

model_a	model_b	mean_exposure_a	mean_exposure_b	wilcoxon_statistic	p_value
biogpt	distilgpt2	2.1937	1.6275	287.5	0.0996
biogpt	gpt2	2.1937	1.7481	252.0	0.0857
distilgpt2	gpt2	1.6275	1.7481	354.0	0.6154

Fig. 7. Pairwise Wilcoxon exposure comparison among the three backbones. The direction of the effect consistently favors higher exposure for BioGPT, although the current sample size does not yield sub-0.05 p-values and the pairwise differences should therefore be interpreted cautiously.

The confidence-interval figure (Fig. 6) strengthens the interpretation by showing that the membership-inference advantage of BioGPT is not marginal. Its interval is compressed near one, while the smaller backbones remain clearly lower. The pairwise-exposure figure (Fig. 7) adds a useful nuance: although the direction of the effect consistently favors higher exposure for BioGPT, the present sample size limits formal pairwise significance. That nuance is important because it keeps the interpretation statistically disciplined while preserving the descriptive cross-metric ordering.

VI. DISCUSSION

The two experiments jointly support several main conclusions.

A. Better Utility Does Not Imply Better Privacy

Experiment I showed that early stopping improved utility substantially but did not reduce leakage. In fact, it produced the highest exposure and strongest membership-inference signal among the three training settings. This is an important practical lesson: interventions typically associated with better generalization in task performance should not automatically be described as privacy defenses. In a clinical LLM setting, the privacy and utility relationship is more nuanced.

At the same time, the combined regularized profile should be interpreted carefully. Because lower learning rate, higher weight decay, and lower-layer freezing were applied together, the present results support the privacy value of that conservative training profile as a whole, but they do not isolate the contribution of each component.

B. Backbone Choice Is a Privacy-Relevant Design Decision

Experiment II yielded a monotonic descriptive ordering across the evaluated backbones. DistilGPT2 had the weakest utility but the lowest privacy risk. GPT-2 improved utility and leaked more. BioGPT, the strongest and most domain-specialized model, achieved the best clinical language modeling but also the highest leakage, exposure, and membership-inference susceptibility. The exposure-based pairwise tests did not reach conventional significance thresholds at the current sample size, but the cross-metric pattern remained consistent across prompt leakage, exposure, and membership inference. Backbone selection in healthcare should therefore be treated not only as a performance decision but also as a privacy-governance decision.

C. Why Did BioGPT Leak More?

The most plausible explanation is not simply model size, but the combination of size and domain specialization. BioGPT began closer to the target distribution and therefore fit the clinical corpus more effectively. Better fit can improve generalization, but it can also increase the probability that the model assigns high confidence to rare, repeated strings such as synthetic canaries. The exposure and membership results support this interpretation. The model was not only better at generating clinically plausible text; it was also more sensitive to whether a canary note had been seen during training.

D. What the Prompt Results Mean

The prompt results suggest that privacy leakage is more easily elicited in clinically natural completion settings than in explicit question-answer forms. Direct recall prompts were generally weak, whereas partial-note completion and prefix-style prompts were more successful. This is important for real-world security because an attacker is not required to ask bluntly for the secret; instead, leakage may surface through

routine-looking note continuations, chart lookups, or audit-style prompts.

E. Interpreting the Saturated Identifier Metric

One metric deserves careful interpretation: patient-identifier recovery rate was 1.0 in several tables. This does not mean the models invented the identifier. The prompts themselves already contained the patient identifier. Accordingly, that metric is less informative than exact-match recovery, condition recovery, per-canary leakage rate, exposure, and membership ROC-AUC. These latter metrics should be emphasized in any formal presentation of the results.

F. Implications for Privacy-Aware Clinical NLP

Taken together, the findings imply that healthcare organizations adapting language models to internal notes should not rely on utility validation alone. A model with excellent perplexity may still leak or encode member-specific traces strongly enough to support membership inference. Conversely, stronger empirical mitigation profiles can reduce leakage, but may impose a non-trivial utility cost. This underscores the need for explicit privacy auditing, careful dataset governance, privacy-preserving representation design [40], and, where feasible, formal defenses such as differential privacy when the underlying data are truly sensitive [36], [37], [39]. The present study is empirical rather than formally private: no differentially private optimization arm was included, and the reported mitigations should not be interpreted as providing a formal privacy guarantee.

G. Practical Recommendations for Deployment

The results support several concrete recommendations. First, privacy auditing should be integrated into model-selection workflows before deployment approval rather than treated as a retrospective security test. Second, backbone choice should be documented in governance reviews as a privacy-sensitive design decision, especially when moving from a smaller general-purpose backbone to a larger or biomedical-specialized one. Third, when private optimization is not yet deployed, conservative regularization profiles and dataset deduplication should be treated as default baselines rather than optional extras. Fourth, prompt-space stress testing should emphasize clinically natural completions and multiple decoding modes, because these were more effective than direct recall prompts in both notebooks. Fifth, deployment review should weigh computational cost together with utility and privacy: in Experiment II, the highest-risk backbone was also the most computationally expensive, so more computation did not buy greater privacy. Finally, institutions fine-tuning on sensitive records should view the present protocol as a minimum empirical audit layer that complements, but does not replace, stronger controls such as secure infrastructure, access governance, red-team testing, data-minimization practice, documented approval workflows, and broader privacy-preserving machine-learning governance frameworks [42].

VII. Scope and Generalization Considerations

The present study is intentionally focused and controlled, which supports clear interpretation of the observed effects. Several points nevertheless help frame the scope of the findings.

1) Corpus scope: The experiments target a multi-specialty clinical transcription corpus with long-form narrative structure. This is a strong setting for comparative privacy analysis, although MTSamples is not a modern production EHR repository and does not fully reflect current documentation workflows, institution-specific templates, longitudinal note chains, or heterogeneous authoring practices. Exact leakage magnitudes may therefore vary for other note distributions, longer contexts, or operational EHR settings.

2) Configuration sensitivity: The reported rankings were obtained under one controlled configuration per experiment, including fixed sequence length, batch size, warmup, optimizer settings, and training duration. The controlled setup supports cleaner comparison, but it does not establish invariance of the results under all plausible hyperparameter changes.

3) Run-to-run uncertainty: The principal comparisons were executed with fixed seeds to preserve strict comparability, and bootstrap confidence intervals were reported for key privacy metrics. Additional repeated-seed experiments would refine uncertainty estimates and runtime-variance analysis without changing the central comparative design.

4) Threat-model coverage: Canary insertion, exposure scoring, and membership inference provide complementary views of memorization. They do not exhaust all possible privacy risks, such as attribute inference, model inversion, gradient leakage, embedding reconstruction, adaptive adversarial prompting, or linkage-based attacks, but they capture three established and measurable leakage pathways in generative-model auditing.

5) Defense scope: The study emphasizes controlled comparison among practical fine-tuning profiles and backbone choices. Formal privacy mechanisms, including end-to-end differentially private optimization, remain an important extension for settings in which auditable privacy guarantees are required.

VIII. Conclusion

A controlled study of privacy leakage and memorization in fine-tuned clinical language models was presented using clinical narrative transcriptions, train-only synthetic canaries, and three complementary attack families. Two experiments were performed. The first compared training settings and showed that early stopping improved utility but did not mitigate privacy risk, whereas a combined regularized training profile eliminated observed prompt leakage and sharply reduced membership-inference susceptibility at the cost of degraded utility. The second compared backbones and showed a clear privacy and utility trade-off: DistilGPT2 was the safest but weakest model, GPT-2 was intermediate, and BioGPT achieved the best utility but also the highest prompt leakage, exposure, and membership-inference risk under the evaluated attacks.

The results support a practical but important conclusion for healthcare AI: *better clinical language modeling can coincide with worse privacy behavior, and backbone choice itself is a privacy-relevant design decision*. For deployable clinical LLMs, privacy auditing should therefore be treated as a first-class evaluation axis alongside utility, latency, and resource

cost. The exposure-based pairwise statistics in Experiment II remained underpowered for conventional significance claims, but the overall descriptive pattern across multiple privacy metrics remained consistent.

Future work should extend the same protocol to credentialed EHR corpora, repeated seeds, formally private optimization, and more adaptive extraction attacks. Nonetheless, the present study already offers a reproducible, defensible, and methodologically strong baseline for auditing privacy leakage in clinical language-model fine-tuning.

ACKNOWLEDGMENT

This research received no external funding. Appreciation is extended to Mohammed First University for providing the computational environment used to conduct the experiments.

REFERENCES

[1] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, R. Saeed, P. N. Ossorio, S. Thadaney-Israni, and A. Goldenberg, "Do no harm: a roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, no. 9, pp. 1337–1340, 2019, doi: 10.1038/s41591-019-0548-6.

[2] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Medicine*, vol. 25, no. 1, pp. 30–36, 2019, doi: 10.1038/s41591-018-0307-0.

[3] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[4] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023, doi: 10.1038/s41591-023-02448-8.

[5] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. T. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[6] L. Y. Jiang, X. C. Liu, N. Pour Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, M. A. Eaton, H. W. Riina, M. J. Laufer, P. S. Kim *et al.*, "Health system-scale language models are all-purpose prediction engines," *Nature*, vol. 619, pp. 357–362, 2023, doi: 10.1038/s41586-023-06160-y.

[7] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020, doi: 10.1038/s42256-020-0186-1.

[8] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Scientific Reports*, vol. 12, art. no. 3034, 2022, doi: 10.1038/s41598-022-05539-7.

[9] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in *28th USENIX Security Symposium*, 2019, pp. 267–284. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[10] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson *et al.*, "Extracting Training Data from Large Language Models," in *30th USENIX Security Symposium*, 2021, pp. 2633–2650. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy*, 2017, pp. 3–18, doi: 10.1109/SP.2017.41.

[12] J. Mattern, F. Mireshghallah, Z. Jin, B. Schölkopf, M. Sachan, and T. Berg-Kirkpatrick, "Membership Inference Attacks against Language Models via Neighbourhood Comparison," in *Findings of ACL 2023*, pp. 11330–11343, 2023, doi: 10.18653/v1/2023.findings-acl.719.

[13] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership Inference Attacks From First Principles," in *2022 IEEE Symposium on Security and Privacy*, 2022, pp. 1897–1914, doi: 10.1109/SP46214.2022.9833649.

[14] S. Ishihara, "Training Data Extraction From Pre-trained Language Models: A Survey," in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, 2023, pp. 171–183, doi: 10.18653/v1/2023.trustnlp-1.23.

[15] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating Training Data Makes Language Models Better," in *Proceedings of ACL 2022*, pp. 8424–8445, 2022, doi: 10.18653/v1/2022.acl-long.577.

[16] F. Mireshghallah, A. Uniyal, T. Wang, D. Evans, and T. Berg-Kirkpatrick, "An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models," in *Proceedings of EMNLP 2022*, pp. 1816–1831, 2022, doi: 10.18653/v1/2022.emnlp-main.119.

[17] Q. Anthony, S. Biderman, U. Prashanth, S. Purohit, E. Raff, H. Schoelkopf, and V. Veliche, "Emergent and Predictable Memorization in Large Language Models," in *Advances in Neural Information Processing Systems 36*, 2023, doi: 10.52202/075280-1219.

[18] A. Aghajanyan, A. Markosyan, K. Tirumala, and L. Zettlemoyer, "Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models," in *Advances in Neural Information Processing Systems 35*, 2022, doi: 10.52202/068431-2773.

[19] S. Ishihara and H. Takahashi, "Quantifying Memorization and Detecting Training Data of Pre-trained Language Models using Japanese Newspaper," in *Proceedings of the 17th International Natural Language Generation Conference*, 2024, pp. 177–188, doi: 10.18653/v1/2024.inlg-main.14.

[20] S. Zhang, H. Li, and R. Ji, "Code Membership Inference for Detecting Unauthorized Data Use in Code Pre-trained Language Models," in *Findings of EMNLP 2024*, pp. 10629–10642, 2024, doi: 10.18653/v1/2024.findings-emnlp.621.

[21] H. Chang, A. S. Shamsabadi, K. Katevas, H. Haddadi, and R. Shokri, "Context-Aware Membership Inference Attacks against Pre-trained Large Language Models," in *Proceedings of EMNLP 2025*, 2025, doi: 10.18653/v1/2025.emnlp-main.370.

[22] D. Parii, T. van Osch, and C. Sun, "Machine Unlearning of Personally Identifiable Information in Large Language Models," in *Proceedings of the Natural Legal Language Processing Workshop 2025*, 2025, doi: 10.18653/v1/2025.nllp-1.6.

[23] K. Edemacu and X. Wu, "Privacy Preserving Prompt Engineering: A Survey," *ACM Computing Surveys*, 2025, doi: 10.1145/3729219.

[24] C. Peris, C. Dupuy, J. Majmudar, R. Parikh, S. Smaili, R. Zemel, R. Gupta, and A. Kumar, "Privacy in the Time of Language Models," in *Proceedings of WSDM 2023*, pp. 1291–1299, 2023, doi: 10.1145/3539597.3575792.

[25] K. Chen, X. Zhou, Y. Lin, S. Feng, L. Shen, P. Wu, and Z. Liu, "A survey on privacy risks and protection in large language models," *Journal of King Saud University Computer and Information Sciences*, 2025, doi: 10.1007/s44443-025-00177-1.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[27] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, art. 2, 2022, doi: 10.1145/3458754.

[28] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022, doi: 10.1093/bib/bbac409.

[29] X. Yang, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. B. Flores, Y. Zhang, T. Magoc *et al.*, "GatorTron: A Large Language Model for Clinical Natural Language Processing," *medRxiv*, 2022, doi: 10.1101/2022.02.27.22271257.

[30] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digital Medicine*, vol. 4, art. 86, 2021, doi: 10.1038/s41746-021-00455-y.

[31] Y. Li, S. Rao, J. R. Ayala Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, A. Rahimi, K. Salimi-Khorshidi, G. Zottoli *et al.*, "BEHRT: Transformer for Electronic Health Records," *Scientific Reports*, vol. 10, art. 7155, 2020, doi: 10.1038/s41598-020-62922-y.

[32] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M. Steinborn *et al.*, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 473–484, 2021, doi: 10.1038/s42256-021-00337-8.

[33] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, art. 119, 2020, doi: 10.1038/s41746-020-00323-1.

[34] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Medical Image Analysis*, vol. 65, art. 101765, 2020, doi: 10.1016/j.media.2020.101765.

[35] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1–19, 2021, doi: 10.1007/s41666-020-00082-4.

[36] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[37] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, and L. Wang, "Differentially Private Fine-tuning of Language Models," *Journal of Privacy and Confidentiality*, vol. 14, no. 1, 2024, doi: 10.29012/jpc.880.

[38] A. El Ouadrhiri and A. Abdelhadi, "Differential Privacy for Deep and Federated Learning: A Survey," *IEEE Access*, vol. 10, pp. 22359–22380, 2022, doi: 10.1109/ACCESS.2022.3151670.

[39] K. Pan, Y.-S. Ong, M. Gong, H. Li, A. K. Qin, and Y. Gao, "Differential privacy in deep learning: A literature survey," *Neurocomputing*, vol. 604, art. 127663, 2024, doi: 10.1016/j.neucom.2024.127663.

[40] Y. Li, T. Baldwin, and T. Cohn, "Towards Robust and Privacy-preserving Text Representations," in *Proceedings of ACL 2018, Short Papers*, 2018, pp. 25–30, doi: 10.18653/v1/P18-2005.

[41] Y. Zhang, H. Pei, S. Zhen, Q. Li, and F. Liang, "Chat Generative Pre-Trained Transformer (ChatGPT) usage in healthcare," *Gastroenterology & Endoscopy*, vol. 2, no. 3, pp. 337–343, 2023, doi: 10.1016/j.gande.2023.07.002.

[42] L. Ramachandrappa, K. H. Krishnappa, A. Salim, R. Warren, R. Madhura, and R. Manasa, "Privacy-Preserving Machine Learning in Healthcare," in *BioMed Bots Assurance*, Boca Raton, FL, USA: CRC Press, 2025, ch. 15, doi: 10.1201/9781003561309-15.

AUTHORS' PROFILE

Yassine Chahid received the M.Sc. degree in science and technology from the Faculty of Science and Technology, Settat, Morocco, in 2014, and the Ph.D. degree in mathematics and computer science from the Faculty of Sciences, Mohammed First University, Oujda, Morocco, in 2022. Since 2014, he has worked with several private-sector companies, where he has held technical and research-oriented positions. He is currently a Technical Lead specializing in artificial intelligence and cybersecurity solutions. His research interests include federated learning, information security, cryptography, and distributed systems.

Anas Chahid received the Engineering degree in computer science from the National School of Applied Sciences (ENSA), Oujda, Morocco. He is currently a Software Engineer and a Ph.D. student in medical artificial intelligence. His work focuses on the design of intelligent and customized software systems. His professional interests include artificial intelligence, data analytics, software engineering, and healthcare applications.

Ismail Chahid received the DUT degree in information technology from the École Supérieure de Technologie, Oujda, Morocco, in 2008, the B.Sc. degree in IT management from the Faculty of Polydisciplinary Studies, Tétouan, Morocco, in 2009, and the M.Sc. degree in business intelligence from the Faculty of Science and Technology, Béni Mellal, Morocco, in 2012. He is currently the Head of the IT Department at the Faculty of Medicine and Pharmacy of Oujda, where he has been working since 2019. His professional interests include business intelligence, data warehousing, information systems management, and software engineering.

Aissa Kerkour El Miad received the M.Sc. degree in operational research and informatics and the Ph.D. degree in computer science from Mohammed First University, Oujda, Morocco. He is currently a Professor with the Department of Computer Science, Faculty of Sciences, Mohammed First University, Oujda. His research interests include image processing, artificial intelligence, high-performance computing, and data mining.

Deployer-Side XAI Instrumentation for Regulated AI: A Clinical Case Study in ICL Sizing

Dorleta Urrutia-Onate[1], Enrique Onieva[2], Asier Perallos[3]
University of Deusto, Avda. Universidades 24, 48007 Bilbao, Spain[1,2,3]
Faculty of Business, Mondragon University, Ibarra Zelaia 2, 20560 Oñati, Spain[1]

***Abstract*—Regulated AI creates a monitoring problem for deployers who must organise human oversight, log-retention and post-market surveillance while often having access only to the prediction interface. This study specifies a deployer-side XAI instrumentation protocol for the output→action boundary, where a model output becomes a reason for action. The protocol reorganises KernelSHAP, nearest-neighbour envelope checks, bounded perturbation, and rank-order stability into a per-decision evidence record computed from `predict()` calls. We instantiate the protocol in a clinical case of phakic Implantable Collamer Lens sizing, using a 55-eye held-out cohort and an Extra Trees regressor for post-operative vault prediction. The record contains five signals: `score_margin`, `constraint_enforcement`, `envelope_validity`, `decision_robustness`, and `record_integrity`, plus two cohort-level oversight aggregates. The case study shows how the same record can support decision-time human oversight, later audit and post-market surveillance under the EU AI Act and the Medical Devices Regulation.**

***Keywords*—Explainable AI; XAI instrumentation; human oversight; Medical Device Software; EU AI Act; ICL sizing**

I. Introduction

Regulated AI creates a practical problem for deployers. They must monitor systems they often cannot inspect. Source code, gradients, training data, weights, tree structures and provider-side documentation may remain unavailable because of proprietary, contractual or organisational restrictions. In many deployments, the deployer can only call the prediction interface. This limited-access setting has to be reconciled with legal duties of human oversight, monitoring, log retention and post-market surveillance under the EU AI Act [1] and the Medical Devices Regulation (MDR) [2]. Explainable AI (XAI) methods can operate from this interface: KernelSHAP [3] against a fixed background sample, distance-based envelope checks and bounded perturbation tests. What is still missing is a specification of which method should produce which evidence, for which deployer-side function, and in what computable form when a prediction is used as a reason for action.

This study specifies that mapping. Its objective is to make a single prediction usable as defensible evidence for action under regulation, using only the prediction interface. The study defines the output→action boundary—the point at which a prediction becomes a reason for action—as its unit of analysis, maps four documented failure modes at that boundary to measurement requirements, and specifies a decision-boundary instrumentation protocol that computes the required evidence from `predict()` calls alone.

This study tests that protocol in a clinical case. Phakic Implantable Collamer Lens (ICL) sizing is a pre-operative decision in which the surgeon chooses one of four available lens sizes. The main safety variable is the post-operative vault, which should remain within a clinically acceptable band to avoid contact- and pressure-related complications [4]. The band used in this study is defined in Section IV-A. A model that predicts post-operative vault from pre-operative biometry and anterior-segment OCT can therefore influence a high-stakes clinical decision. In the regulatory setting used here, the tool is treated as Medical Device Software under the MDR and as high-risk Medical Device Artificial Intelligence under Annex III of the AI Act. The clinical deployer is then responsible for organising oversight in use.

The protocol emits a per-decision telemetry record whose five core signals are `score_margin`, `constraint_enforcement`, `envelope_validity`, `decision_robustness` and `record_integrity`, complemented by two cohort-level oversight fields, `sustainability_impact` and `human_ai_teaming`. It is instantiated end-to-end on a 55-eye held-out cohort of an ICL vault prediction regressor (Extra Trees, held-out mean absolute error 160.37 µm), producing per-decision JSON records and a 4×55 counterfactual matrix over the lens-size decision space, and the resulting record is mapped to EU AI Act Article 14 and MDR Articles 83 and 88. Model performance and lens-size superiority are outside the study's claim.

The study is organised as follows: Section II reviews XAI methods, the deployer's position under regulation, the standards frame and the clinical context of ICL sizing. Section III defines the methodological framework and the output→action boundary as the unit of analysis. Section IV introduces the case study, dataset, deployment context and deployed regressor. Section V develops the signal architecture under limited-inspection deployment. Section VI applies the protocol to the held-out cohort, reports the per-decision records, estimates cohort-level review workload and analyse the counterfactual sweep over the four lens sizes. Section VII and Section VIII discuss the findings and conclude.

II. Background and Related Work

A. XAI Methods and their Classical Destination

Post-hoc XAI is usually grouped into four method families. Feature-attribution methods, including KernelSHAP [3] and LIME [5], assign a per-feature contribution to a single prediction. Counterfactual-explanation methods [6], including DiCE-style procedures, search the input neighbourhood for changes

that would alter the model output. Prototype- and influence-based methods locate a prediction in relation to training examples. Surrogate-tree methods approximate an opaque model with a globally interpretable structure. Arrieta et al. [7] provide the taxonomy used here as background.

Recent work has made the model-centred use of post-hoc XAI harder to defend without qualification. Adebayo et al. [8] showed that several saliency methods can produce visually similar attributions even after model randomisation. Slack et al. [9] showed that LIME and SHAP can be manipulated by an adversary controlling the prediction interface, leaving the explanation blind to fairness-relevant features. Rudin [10] argued that high-stakes decisions should use interpretable models where possible instead of post-hoc explanations for opaque models.

A second critique matters more for deployment: who is meant to use the explanation, and for what. Bhatt et al. [11] document that XAI outputs in many organisations are still mainly consumed by model developers, rather than by the operators who must act on model outputs. Liao and Varshney [12] argue that XAI remains incomplete when the user question is not specified. Doshi-Velez and Kim [13] anticipated this separation by distinguishing functionally grounded proxies from human-grounded and application-grounded evaluations.

The literature, therefore, separates two uses of post-hoc XAI. One use serves development: model construction, debugging and selection. It works close to the model and characterises model-level behaviour. The other use serves deployment: a specific output is about to be used as a reason for action. It works from the prediction interface and must characterise one decision in context.

Table I summarises this distinction. Fig. 1 locates the point where the two uses meet: the output→action boundary. Most XAI methods, including the methods used in this study, were designed for model-centred work. The question here is how to reorganise them into deployer-side evidence records.

For the deployer, the useful destination of XAI is the decision record. A SHAP attribution, a counterfactual explanation or a surrogate model is useful in regulated deployment only if it helps document why a specific output could be used, disregarded, overridden or reviewed at that point in the workflow. This study, therefore, treats XAI as instrumentation for functional characterisation, observable behaviour analysis, validation and technical trust evidence at decision time.

B. The Deployer's Predicament Under Regulation

The EU AI Act distinguishes the provider that develops or places an AI system on the market from the deployer that uses the system under its authority. Article 26 assigns deployers of high-risk AI systems duties around use according to instructions, assignment of human oversight, monitoring of operation, reporting of risks or serious incidents, and retention of automatically generated logs when those logs are under their control [1]. Article 14 defines the human-oversight capabilities that high-risk systems must enable: understanding capacities and limitations, remaining aware of automation bias, interpreting outputs and deciding when to disregard, override or reverse an output [1].

The MDR adds clinical evaluation, post-market surveillance and vigilance [2]. For AI-enabled medical devices, the MDCG–AIB guidance [14] clarifies that the AI Act and the MDR/IVDR apply together, while separating the AI Act deployer from the MDR manufacturer. In many clinical uses, the surgeon, hospital or clinic operating the tool is the deployer.

This creates a gap between responsibility and access. The deployer must monitor a system in use, while model internals, training data, source code and provider-side validation material may remain unavailable for technical, legal, organisational or commercial reasons [15], [16], [17]. This study addresses that access regime: the deployer has the case input, a fixed background sample and the ability to call `predict()`.

C. Regulatory and Standards Frame

The operational frame combines the binding regulation of Section II-B with voluntary management standards.

Voluntary frameworks give organisations procedures for organising AI risk and evidence. The NIST AI Risk Management Framework 1.0 [18] organises that work through four functions: *Govern*, *Map*, *Measure* and *Manage*. ISO/IEC 42001:2023 [19] specifies an AI management system based on a plan-do-check-act cycle and auditable operational records.

These frameworks define the organisational need: risks must be governed, mapped, measured, managed and recorded. The protocol developed here supplies an operational mechanism for producing such records at the point where an AI output may become a reason for action.

D. Clinical Context: ICL Sizing as a Regulated AI Decision

The case study uses phakic ICL sizing for refractive correction. The ICL outcomes literature commonly treats the band $[250, 750]$ µm as the range of clinical sufficiency for post-operative vault [4]. This study adopts that band as the operational decision boundary for the case study (Section IV-A).

An ICL sizing tool can influence a high-stakes clinical decision. The model output can shape the lens size selected before surgery. If the predicted vault lies outside the clinically acceptable band, the output becomes part of the reasoning behind a surgical decision with safety implications. Under the medical-device regime, that use needs evidence for oversight, validation and post-market monitoring.

Clinical XAI literature raises the same concern from the healthcare side. Babic et al. [20] warn that explanations in healthcare can overstate what they show, especially when post-hoc rationalisations do not reflect the basis of the prediction. Tonekaboni et al. [21] show that clinicians ask for explanations tied to the clinical situation, timing and decision context, while many XAI methods provide outputs that do not match those needs. Amann et al. [22] frame explainability in healthcare as a multidisciplinary requirement linked to ethical values, safe use and accountability. These contributions support the move from model-centred explanation to decision-time evidence.

The dataset used here comes from clinical operations at three centres of a multi-site ophthalmology network. It was first organised for predictive modelling in earlier clinically

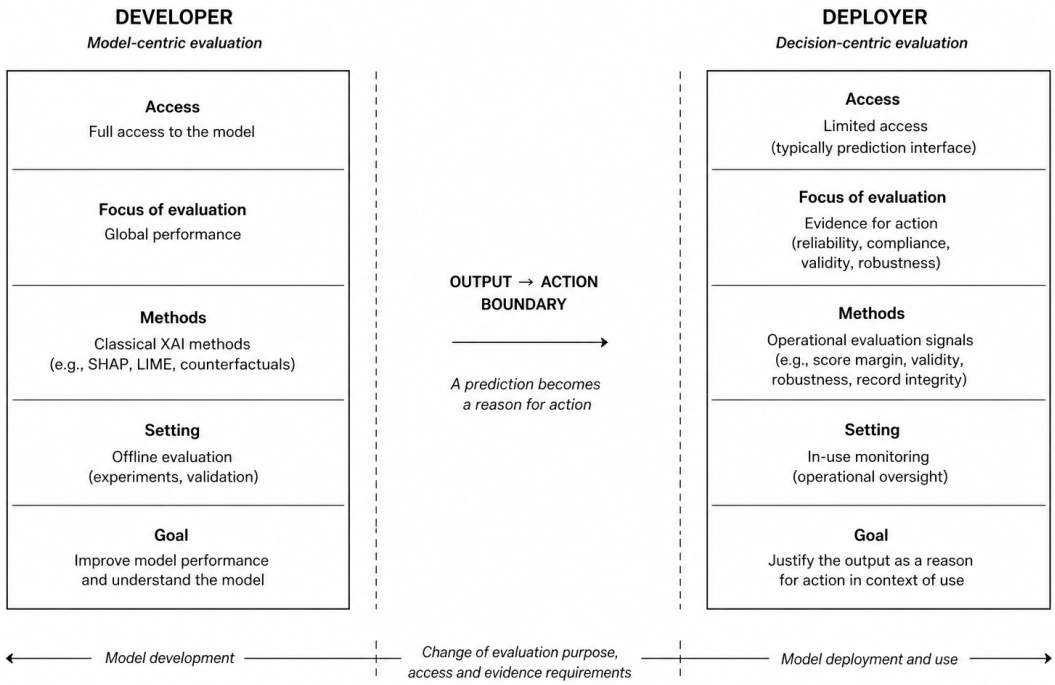


Fig. 1. Developer-side and deployer-side destinations for post-hoc XAI under regulated deployment. The developer-side destination is model-centred: it assumes broad access to the model and supports development, debugging and model selection. The deployer-side destination is decision-centred: it usually works from the prediction interface and supports evidence for action, in-use monitoring and audit. The two destinations meet at the output→action boundary, where a prediction becomes a reason for action.

TABLE I. DEVELOPER-SIDE XAI VS. DEPLOYER-SIDE INSTRUMENTATION: TWO DESTINATIONS FOR POST-HOC EXPLANATION UNDER REGULATED DEPLOYMENT.

Aspect	Developer-side XAI	Deployer-side instrumentation
Access to the model	weights, gradients, training data, architecture, internals	prediction interface only (`predict()` calls)
Unit of analysis	the model	the individual decision
Phase of the lifecycle	development, debugging, model selection	deployment, decision time, post-market surveillance
Primary consumer	model developer, ML researcher	deployer: operator, clinician, audit team
Output form	global explanations, fidelity diagnostics, debugging outputs	per-decision evidence record + cohort aggregates
Regulatory anchor	usually indirect	EU AI Act Art. 14 and Art. 26; MDR Art. 83 and Art. 88
Question answered	does this model behave as expected in general?	is this specific output defensible as a reason for action now?

supervised work. This study inherits the clinical motivation, feature inventory and filtering criteria from that work.

The dataset and the regression model are used to expose the AI-decision structure on which the instrumentation protocol is tested.

E. Accountability Artefacts and Decision Logging

A parallel line of work proposes documentation and logging artefacts for accountable AI. Model Cards [23], FactSheets [24] and Datasheets for Datasets [25] standardise what a provider should disclose about a model, a service or a dataset, while internal audit frameworks such as the one of Raji et al. [26] organise accountability as an organisational process. Decision provenance [27] extends this from static documentation towards logging the data flow behind individual decisions. These artefacts are mostly produced at model or dataset release, on the provider or developer side, and they characterise the system rather than a single output. The protocol developed here is complementary and sits at a different point: it is a per-decision, deployer-side evidence record computed at prediction time from `predict()` calls, so that the actor who acts on an output can attach defensible, auditable evidence to that specific decision under limited inspection.

F. Position of this Work

The architecture maps the deployer's monitoring problem to four failure modes at the output→action boundary, each turned into a measurement computable from `predict()` calls alone (Section III). This study adapts it to a clinical regression case and applies it end to end to a real ophthalmology cohort.

The ICL case has three useful properties for the protocol.

First, the predictive task is regression on a continuous safety variable, so the relevant boundary is a clinical band rather than a probability threshold. Second, the dataset comes from a real multi-centre clinical cohort operating under the medical-device regime. Third, the decision space is small and discrete: the surgeon chooses one of four lens sizes. That discreteness allows the counterfactual component to enumerate all decision alternatives. For each held-out eye, the protocol queries the model under each candidate lens size while holding anatomy and biometry fixed, and then recomputes the boundary-relevant signals for each alternative. The premise is narrow. XAI outputs used in regulated deployment must be structured well enough to count as evidence. The question is whether each decision record supports functional characterisation, observable behaviour analysis, validation and technical trust evidence at the moment when a prediction is used as a reason for action.

III. Methods

This section defines the framework applied in the clinical case. The framework was introduced in the companion manuscript [28]. Here it is adapted to a regression task on a continuous safety variable and later implemented for the ICL vault prediction model.

A. Unit of Analysis: The Output→Action Boundary

The framework takes the output→action boundary as its unit of analysis. This is the point at which a single prediction becomes a reason for action in a clinical, financial or industrial workflow.

At that point, the question is evidential. The deployer responsible for monitoring the system in use must be able to attach evidence to the prediction before the decision is taken. Under the access conditions described in Section II-B, that evidence must be computable from the prediction interface. The framework therefore specifies what evidence is needed, how it is structured and which part of the output→action relation it characterises.

B. From Failure Modes to Measurement Requirements

Four safety and human-factors perspectives identify failure modes at the output→action boundary.

Leveson [29] identifies constraint-enforcement failure: the action proceeds on grounds that the institution would not recognise as legitimate. Rasmussen [30] identifies envelope migration: the case has moved beyond the population on which the model was validated. Parasuraman [31] identifies automation bias: a fragile output is treated as robust because the operator has no evidence of fragility. A fourth failure is procedural: accountability failure, where no reproducible record links the output, the decision context and the evidence used at the time of action.

Each failure mode defines a measurement requirement. Constraint-enforcement failure requires evidence about whether the local prediction is supported by legitimate variables or by decision-linked variables. Envelope migration requires evidence about the distance between the case and the validated population. Automation bias requires evidence about whether decision-band membership is stable under bounded perturbation. Accountability failure requires evidence about whether the explanation is stable enough to be used as a record.

The framework also includes a domain-derived signal. In the ICL case, the predicted vault must be located relative to the operational clinical band. This signal indicates whether the predicted vault is safely inside the band, close to a threshold or outside the clinically acceptable range.

C. From Measurement Requirements to Prediction-Interface Methods

The deployer's access condition restricts the available methods. The framework therefore selects XAI and validation methods that can operate from the prediction interface.

KernelSHAP [3] computes feature attributions from `predict()` calls against a fixed background sample. It can therefore measure attribution mass on legitimate and decision-linked variables. *k*-nearest-neighbour distance operates on the input feature space and measures envelope validity. Bounded perturbation repeatedly calls `predict()` on perturbed versions of the same input and estimates the flip rate used for `decision_robustness`. Rank-order stability under micro-noise recomputes the attribution procedure on small perturbations of the input and produces `record_integrity`.

The methodological step is the binding between failure mode, measurement requirement, computable signal and prediction-interface method. Each XAI or validation method receives a specific evidential function at the output→action boundary.

D. Properties of the Resulting Protocol

The protocol has three methodological properties.

First, the protocol is compatible with prediction-interface-only access. It requires the case input, a fixed background sample and the ability to call `predict()`. It does not require model weights, gradients, trees, source code, training data or provider-side documentation.

Second, the protocol is non-compensatory. Each signal addresses a different failure mode. A high value on one signal does not offset a failure on another. A stable attribution does not compensate for a prediction outside the clinical band. Robust band membership does not compensate for reliance on decision-linked variables.

Third, the protocol is decision-boundary local. Each signal is computed for the specific case on which the deployer is acting. The protocol produces a per-decision evidence record attached to one prediction at the moment that prediction may support action. Model-level validation remains necessary, but it answers a different question.

IV. Case Study

A. Clinical Target and Decision Space

The clinical target of an ICL sizing model is the post-operative vault: the distance between the implanted lens and the natural crystalline lens. A vault below the clinically acceptable range may increase the risk of contact-related

complications. A vault above that range may increase intraocular pressure and the risk of secondary glaucoma. The ICL outcomes literature commonly uses $[250, 750]$ µm as the range of clinical sufficiency [4].

The decision supported by the tool is the pre-operative choice of lens size. In this case, the available sizes are the four standard options used at the contributing sites: {12.1, 12.6, 13.2, 13.7} mm. The decision uses anterior-segment optical coherence tomography (OCT), biometric measurements and the clinician's selected refractive correction.

The regression model analysed here evaluates candidate decisions. For each candidate lens size and selected refractive correction, the model uses the available anatomical and biometric measurements to predict the post-operative vault expected from that combination. The model supports sizing by estimating the safety variable associated with each candidate decision.

B. Dataset

The dataset comes from clinical ophthalmology operations at three centres of a multi-site ophthalmology network. The data were first organised for predictive modelling in earlier clinically supervised work. This study inherits three elements from that work: clinical motivation, feature inventory and cohort filtering criteria.

The clinical motivation is vault prediction, as the principal safety variable in ICL sizing. The feature inventory contains 25 pre-operative variables combining anterior-segment OCT indices, biometric and refractive measurements, and categorical descriptors of the implant decision. The filtering criteria retain eyes with complete pre-operative OCT, biometry and post-operative vault measurement. After filtering, the cohort contains 545 eyes.

The dataset is reused to evaluate whether the instrumentation protocol can produce decision-level evidence from a deployed regressor.

The feature set contains 25 pre-operative variables grouped into three sets.

The OCT set includes anterior-segment indices: lens rise, anterior depth, limbus diameter and centre, pupil diameter and centre, and the temporal/nasal AOD/TISA pairs at the 750 µm reference.

The biometric and refractive set includes k1 keratometry, patient age at operation, pre-operative spherical equivalent and selected refractive correction.

The decision-descriptor set includes categorical variables encoded as one-hot indicators: lens family (`ICL`, `TICL`), refractive type (`Myopic`, `Toric Myopic`) and lens-size dummies (`lens_12.1`, `_12.6`, `_13.2`, `_13.7`).

The lens-size variables are encoded as four independent indicators, not as one numerical variable. This avoids treating lens size as a continuous ordered scale. It also aligns the feature representation with the discrete decision space defined in Section IV-A. The same encoding supports the counterfactual sweep in Section VI-C, where each eye is evaluated under the four candidate lens sizes.

All 25 features are available at the moment of the sizing decision. This matters for the deployment-time evaluation in Section VI: the model is evaluated using only information available before the clinical decision. Continuous features are imputed with the median of the training partition and the same imputers are applied to the held-out partition, avoiding held-out information in the imputation step.

Case identifiers used in the body of the study are anonymised as `case1`, `case2`, ..., `case5`. These labels do not correspond to clinical-record database entries.

C. Deployment Context

ICL sizing models are no longer only isolated research prototypes. Since 2021, several machine-learning pipelines have been published for vault prediction and lens-size selection on cohorts ranging from hundreds to thousands of eyes. These include Random Forest and gradient-boosted regressors trained on large EVO-ICL records [32], regression-tree ensembles trained on multicentre OCT–MS39 data [33], and deep-learning models trained on raw SS-OCT images and made publicly available through the `safevaulticl.com` interface [34].

The model studied here belongs to that emerging class of clinical decision-support software. Its relevance for this study is the operational structure of the decision: a model output can become part of the reasoning behind a surgical choice.

Section II set out the regulatory frame. In this case, the relevant actor is the clinical user of the system. In a clinical setting, the operator of the sizing AI is typically the refractive surgeon or the clinic's data team, not the model developer. This actor is the deployer under Article 26 of the EU AI Act and must organise system use, monitoring and human oversight in practice [1].

Article 14 defines the human-oversight capacities that the system must enable. Article 26 places the deployer in the operational position where those capacities must be used. MDR Articles 83 and 88 add the post-market surveillance and trend-reporting frame for the medical-device setting [2].

The problem is operational. The deployer is expected to monitor the system in use, while access to the model's internal design, training process and provider-side documentation may be limited [15], [16], [17]. The case study asks what evidence the deployer can document when an ICL vault prediction is used as a reason for a sizing decision.

D. The Deployed Regressor

From the 545-eye dataset described in Section IV-B, 490 eyes are used for training and 55 for held-out evaluation. The split is reproducible, with `test_size = 0.10` and `seed = 42`.

Five standard scikit-learn regressors are evaluated on the training partition using 10-fold cross-validation and mean absolute error (MAE) as the scoring metric: Extra Trees [35], Random Forest [36], Ridge regression, histogram-based gradient boosting and standard gradient boosting [37]. Table II reports the cross-validated MAE for each candidate.

TABLE II. CROSS-VALIDATED COMPARISON OF CANDIDATE REGRESSORS ON THE TRAINING PARTITION (490 EYES, 10-FOLD CV).

Rank	Model	MAE_cv (μm)
1	**Extra Trees** *(selected)*	**167.24**
2	Ridge	175.73
3	Random Forest	176.86
4	HistGradientBoosting	180.50
5	Gradient Boosting	182.99

The Extra Trees regressor is selected for the lowest cross-validated MAE; on the held-out partition its MAE is 160.37 μm, close to the cross-validated training value, which we read as evidence of no obvious train/held-out drift rather than a precision claim.

The model is serialised with its column order, per-feature imputation medians and lens-size column name; this object is the only model input consumed by the protocol in Section V. The regressor is not meant to outperform the clinical state-of-the-art: it is the test bed for the instrumentation protocol.

V. SIGNAL ARCHITECTURE UNDER LIMITED-INSPECTION DEPLOYMENT

A. Failure Modes at the Output→Action Boundary and the Measurements they Require

Sections II-B and Section II-C define the deployer-side access condition. Section III defines the failure-mode/measurement-requirement chain. This section turns that chain into operational signals and prediction-interface methods. The clinical case adds one domain signal: the distance between the predicted vault and the centre of the operational clinical band. That domain signal orients the four failure-mode signals. The methodological contribution is the binding itself: existing XAI and validation methods are assigned, one by one, to specific failure modes at the output→action boundary under prediction-interface-only access. Table III summarises the chain.

Each signal follows from a failure mode and from the access conditions of deployment. The chain is: failure mode, required measurement, computable signal, prediction-interface method.

B. Cohort-Level Aggregates: From Per-Decision Telemetry to Oversight Workload

The five signals in Section V-A, together with case identifier, timestamp, model version and lens-size selection, form the telemetry record emitted at the output→action boundary. The operator can use this record at decision time and persist it afterwards as the audit unit.

The record also contains two oversight fields: `sustainability_impact` and `human_ai_teaming`. Their cohort-level aggregates estimate how many outputs remain inside the operational band and how many can proceed under the AI-supported workflow before human review. This matters because Article 14 requires oversight measures that allow assigned persons to understand system capacities and limits, remain aware of automation bias, interpret outputs and decide when to disregard, override or reverse them [1].

`human_ai_teaming_cohort` measures the proportion of decisions that satisfy the joint sufficiency criterion envelope_validity $\geq \tau_{\text{env}}$ and score_margin $\geq \tau_{\text{margin}}$. Its complement is the expected human-review workload under Article 14. Both aggregates are computed from the same per-decision telemetry and require no additional measurements. The thresholds $(\tau_{\text{env}}, \tau_{\text{margin}})$ are set by the deployment context, balancing unnecessary routing against missed review.

C. Pipeline and Per-Decision Evidence Record

The three methodological properties in Section III-D instantiate, as shown in Fig. 2. The Listing given below gives one telemetry record from the held-out cohort. The same record is the decision-time view for the clinician and the later audit unit used by post-market surveillance.

Section VI reports what the protocol produces on every held-out eye.

VI. APPLICATION TO THE ICL VAULT PREDICTION MODEL

This section applies the protocol of [28] to the deployed ICL vault regressor, showing the evidence it produces from `predict()` calls only, for one sizing decision and for cohort-level oversight. Box 1 presents the glossary signals and parameters used in Section VI.

Section VI-A defines the calibration choices; Section VI-B reports the output on the 55-eye held-out cohort; Section VI-C evaluates the four lens sizes and the counterfactual matrix for oversight and post-market surveillance; Section VI-D states reproducibility.

A. Operational Calibration

Applying the architecture (Section V) to the deployed regressor (Section IV-D) requires four calibration choices: the clinical band (`score_margin`), the feature taxonomy (`constraint_enforcement`), the perturbation parameters (`decision_robustness`, `record_integrity`), and the operating point for cohort-level oversight. The clinical band $[250, 750]$ μm is taken from Section IV-A unchanged; Table IV summarises the calibration.

Clinical band. The `score_margin` signal reports the distance of the predicted vault from the centre of the clinical band, normalised by the band half-width. With centre $c = 500$ μm and half-width $h = 250$ μm, the signal is:

$$\text{score_margin}(\hat{y}) = 1 - \frac{|\hat{y} - c|}{h} \tag{1}$$

A prediction at the band centre yields 1.0, at either edge 0.0, and outside the band a negative value whose magnitude indicates how far outside the case lies.

Feature taxonomy. The legitimate/decision-linked split defines the clinical meaning of `constraint_enforcement`. In this case, the split separates variables that describe the eye from variables that encode, or partly encode, the clinical decision. The deployer needs to know whether the local prediction is mainly supported by anatomical and biometric

TABLE III. FAILURE MODES AT THE OUTPUT→ACTION BOUNDARY, THE MEASUREMENT THEY REQUIRE, THE SIGNAL THAT DELIVERS IT, AND THE PREDICTION-INTERFACE-COMPUTABLE METHOD.

Perspective	Failure	Measure	Signal	Method
Domain (clinical band)	Case far from operational threshold	Distance from band centre	`score_margin`	scalar of the prediction
Leveson (STAMP) [29]	Action proceeds on non-legitimate or decision-linked features	Share of attribution mass on legitimate features	`constraint_enforcement`	KernelSHAP [3] vs fixed background
Rasmussen [30]	Case outside validated envelope	Distance to validated population	`envelope_validity`	k-NN distance in feature space
Parasuraman [31]	Brittle output read as robust	Flip-rate under bounded perturbation	`decision_robustness`	repeated `predict()` over perturbations
Cross-cutting accountability	No reproducible record of evidential basis	Rank-order stability of the attribution	`record_integrity`	KernelSHAP recomputed under micro-noise

Listing. Example structure of one per-decision telemetry record.

```
{
  "case_id": "case1",
  "timestamp": "2026-04-30T12:34:56Z",
  "model_version": "extra_trees_icl_v1",
  "predicted_vault_um": 498.5,
  "lens_option_mm": 13.2,
  "core_signals": {
    "score_margin": 0.99,
    "constraint_enforcement": 0.73,
    "envelope_validity": 1.00,
    "decision_robustness": 1.00,
    "record_integrity": 0.86
  },
  "oversight_fields": {
    "sustainability_impact": "inside_clinical_band",
    "human_ai_teaming": "ai_supported_workflow_sufficient"
  },
  "counterfactual_by_lens_mm": {
    "12.1": {"vault_um": 316, "score_margin": 0.27, "in_band": true},
    "12.6": {"vault_um": 351, "score_margin": 0.40, "in_band": true},
    "13.2": {"vault_um": 499, "score_margin": 0.99, "in_band": true},
    "13.7": {"vault_um": 684, "score_margin": 0.26, "in_band": true}
  }
}
```

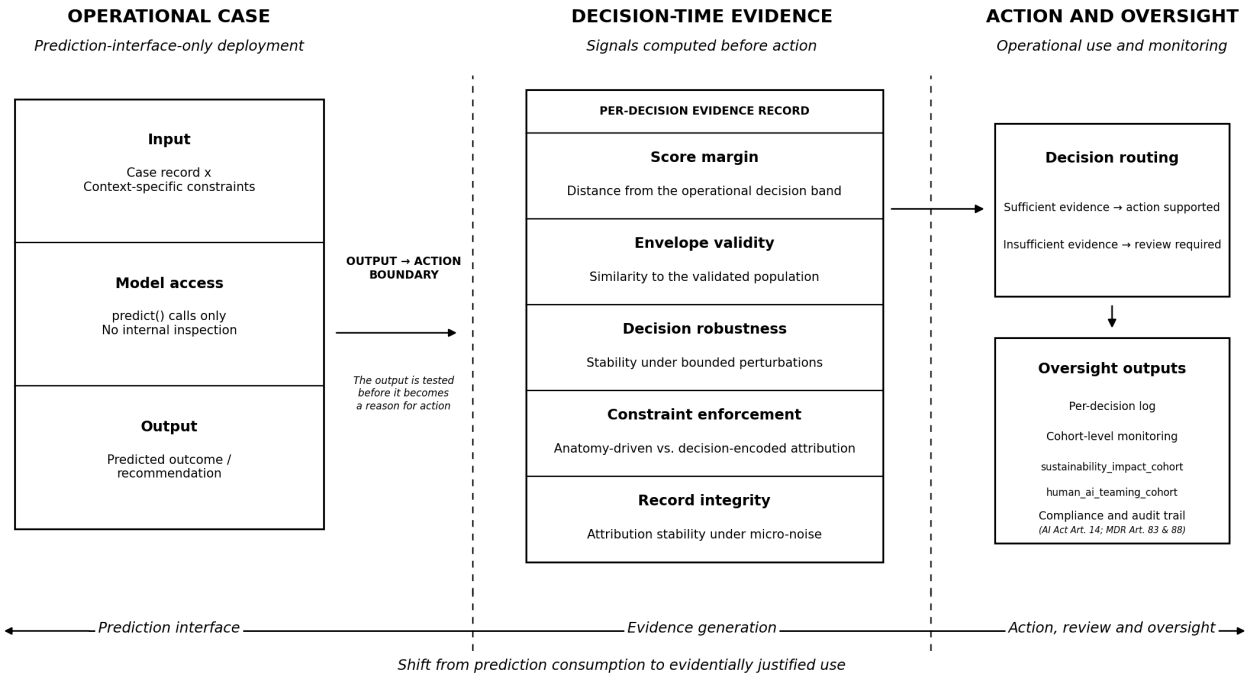


Fig. 2. Pipeline of the protocol on a single decision. The deployer queries the model through `predict()` only. The five core signals and two per-decision oversight fields are emitted as a JSON evidence record. Cohort-level aggregates are computed later by aggregating persisted records across cases.

Box 1 — Glossary of signals and parameters used in Section VI. The five core signals, the two cohort-level oversight aggregates and the calibration parameters used throughout this section are summarised below. Operational definitions and formulas are given in Section VI-A. Cohort-level results are reported in Section VI-B to Section VI-C. Dataset feature names with their legitimate/decision-linked classification appear in Table V.

Symbol / field	Description	Range / value
`score_margin`	distance to clinical-band centre, normalised by half-width	$(-\infty, 1]$
`constraint_enforcement`	KernelSHAP attribution mass on legitimate features	$[0, 1]$
`envelope_validity`	k-NN distance beyond validated population, clipped	$[0, 1]$
`decision_robustness`	fraction of perturbations preserving band membership	$[0, 1]$
`record_integrity`	mean Spearman rank-order stability under micro-noise	$[-1, 1]$
`sustainability_impact_cohort`	cohort fraction with prediction inside [250, 750] µm	$[0, 1]$
`human_ai_teaming_cohort`	cohort fraction satisfying joint sufficiency	$[0, 1]$
τ_{env}	sufficiency threshold for `envelope_validity`	0.5
τ_{margin}	sufficiency threshold for `score_margin`	0.3
σ	bounded-perturbation magnitude for `decision_robustness`	$0.10 \times$ std
ε	micro-noise magnitude for `record_integrity`	$0.01 \times$ std
N	bounded-perturbation samples per case	100
M	micro-noise replicates per case	50
k	k-NN neighbourhood size	5
$d_{\max}$	envelope distance threshold	90th percentile of train 5-NN mean
X_{bg}	KernelSHAP background sample size	50
$\mathcal{L}, \mathcal{D}$	legitimate / decision-linked feature partition (Table V)	14 / 11

Box 2 — Justification of calibration parameters. Each non-domain parameter used in Section VI-A is summarised here with its operational rationale and references.

Parameter	Value	Rationale	Reference / source
Clinical band $[v_{\text{lo}}, v_{\text{hi}}]$	$[250, 750]$ µm	domain-derived; clinical literature on ICL vault outcomes	[4]
X_{bg}	50	background sample size for KernelSHAP; in the recommended 50–1000 range, with empirical stability of SHAP explanations documented for sample sizes of this order	SHAP documentation; [38]
k	5	small neighbourhood size for k-NN-based out-of-distribution detection on a 490-eye reference set; consistent with deep k-NN OOD-detection literature	[39]
$d_{\max}$ rule	90th percentile of train 5-NN mean	percentile-based threshold for out-of-distribution detection; less sensitive to training-set outliers than a maximum-distance rule	[39]
σ	$0.10\times$ std	bounded-perturbation magnitude; of the order of inter-operator anterior-segment OCT measurement variability reported in the imaging literature	domain-derived; calibrated against OCT inter-operator variability
ε	$0.01\times$ std	micro-noise magnitude one order of magnitude below σ; tests rank-order stability of the explanation under noise that should leave clinical interpretation unchanged	derived; clinical-interpretation invariance
N	100	bounded-perturbation samples per case; standard Monte Carlo budget for flip-rate estimates at the precision used by `decision_robustness`	empirical; pilot stability test
M	50	micro-noise replicates per case; sufficient sample for the Spearman rank-order correlation estimate used by `record_integrity`	empirical; pilot stability test
$(\tau_{\text{env}}, \tau_{\text{margin}})$	$(0.5, 0.3)$	deployer-side operating point; conservative defaults used throughout this study. Sensitivity at three alternative values is reported in Table VI	deployer-side domain choice

TABLE IV. OPERATIONAL MAPPING OF THE 5 + 2 SIGNALS ON THE ICL VAULT PREDICTION MODEL

Signal	Domain choice	Parameters
`score_margin`	clinical band [250, 750] µm	half-width $h = 250$, centre $c = 500$
`constraint_enforcement`	L = 14, D = 11 (Table V)	$X_{\text{bg}} = 50$, `seed = 42`
`envelope_validity`	25-D feature space, post-imputation	$k = 5$, $d_{\max} =$ max train 5-NN mean
`decision_robustness`	bounded Gaussian perturbation on continuous features	$N = 100$, $\sigma = 0.10$ std
`record_integrity`	rank-order stability under micro-noise on continuous features	$M = 50$, $\varepsilon = 0.01$ std
`sustainability_impact_cohort`	inside clinical band	population fraction
`human_ai_teaming_cohort`	sufficiency thresholds	$\tau_{\text{env}} = 0.5,\ \tau_{\text{margin}} = 0.3$

measurements, or whether it is reproducing variables derived from the clinician's prospective choice.

The 25 pre-operative features are divided into a legitimate set $\mathcal{L}$ of fourteen anatomical, biometric, age and laterality variables, and a decision-linked/report-dependent set $\mathcal{D}$ of eleven variables. The full taxonomy is shown in Table V.

The legitimate set contains pre-operative anatomical and biometric measurements that exist independently of the sizing decision. The decision-linked/report-dependent set contains features whose values encode the clinician's prospective lens-size selection (`lens_*`, `selected_lens_size`, `selected_sph`), lens-family and refractive-type dummies linked to the clinical decision (`ICL`, `TICL`, `Myopic`, `Toric Myopic`), and patient-reported refraction, which is subject to instrument-of-reporting variability (`preop_sph`).

The *legitimate/decision-linked* partition is informational, not a judgement of clinical validity: lens size physically influences vault, so decision-linked variables are clinically relevant inputs. `constraint_enforcement` asks whether the prediction differentiates cases mainly through anatomical evidence (a high value: anatomy and biometry carry most of the attribution mass) or by tracking the lens choice already encoded in the input (a low value: dominated by decision-linked or report-dependent variables). Both modes can be clinically plausible and provide different evidence at the output→action boundary.

In the counterfactual sweep of Section VI-C, the lens-size variables are queried under each of the four candidate values as intervention variables. By contrast, `constraint_enforcement` is computed at the originally observed prediction, with all input features held at their recorded values.

Two caveats about the KernelSHAP background sample X_{bg}: it is drawn from the training partition here for reproducibility, but in deployment without training-data access the provider should supply an anonymised representative sample, since KernelSHAP needs a reference distribution for per-feature contributions [3]; and because KernelSHAP interpretations depend on that sample [38], the protocol relies on a provider-supplied or training-derived sample the deployer cannot fully verify from the prediction interface alone.

The `constraint_enforcement` signal is computed as:

$$\text{constraint_enforcement} = \frac{\sum_{i\in\mathcal{L}} |\phi_i|}{\sum_{i\in\mathcal{L}} |\phi_i| + \sum_{i\in\mathcal{D}} |\phi_i|}. \quad (2)$$

Here, ϕ_i is the per-feature KernelSHAP attribution computed by the protocol's regression-aware KernelSHAP variant. The cohort-level aggregates reported in Sections VI-B–VI-C are computed from these per-decision records.

1) Validation envelope: The `envelope_validity` signal reports whether the case remains inside the feature-space envelope of the training partition. The signal uses a *k*-nearest-neighbour distance estimator on the 25-D post-imputation feature space, with $k = 5$. Features are standardised with the training-partition mean and standard deviation before distance computation, so each variable contributes comparably to the Euclidean distance.

The threshold $d_{\max}$ is the 90th percentile of the per-training-point 5-NN mean distance over the 490-eye training partition, a choice consistent with out-of-distribution detection practice [39]. The signal is:

$$\text{envelope_validity}(x) = 1 - \max\left(0,\ \frac{d_{k\text{NN}}(x)}{d_{\max}} - 1\right), \quad (3)$$

clipped to $[0, 1]$: a value of 1.0 means the case remains within the calibrated training envelope, and lower values indicate greater distance beyond the threshold.

2) Perturbation parameters: The bounded-perturbation magnitude used by `decision_robustness` is set to $\sigma = 0.10$ of each continuous feature's training-partition standard deviation. Perturbations are applied only to continuous anatomical, biometric and refractive variables. One-hot categorical indicators and the candidate lens-size encoding are held fixed during bounded-perturbation and micro-noise tests.

This magnitude is of the same order as inter-operator OCT measurement variability, so a band-membership flip under it means the decision can change under deployment-level measurement noise.

Let s_j denote the training-partition standard deviation of continuous feature j. With $N = 100$ perturbations $\epsilon_n \sim \mathcal{N}(0, \text{diag}((\sigma s_j)^2))$ and band-membership indicator $b(\hat{y}) = \mathbf{1}[250 \le \hat{y} \le 750]$, the signal is:

$$\text{decision_robustness}(x) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\Big[b(\hat{f}(x+\epsilon_n)) = b(\hat{f}(x))\Big]. \quad (4)$$

A value of 1.0 indicates that no perturbation flips the case's band membership; lower values report the share that do.

The micro-noise magnitude used by `record_integrity` is set to $\varepsilon = 0.01$ of each continuous feature's training-partition standard deviation. This tighter regime tests whether the attribution ranking stays stable under noise that should leave the clinical interpretation unchanged.

With $M = 50$ micro-perturbed copies $\delta_m \sim \mathcal{N}(0, \text{diag}((\varepsilon s_j)^2))$ and Spearman rank-order correlation ρ_S, the signal is the mean correlation between the original attribution vector and each perturbed replicate:

$$\text{record_integrity}(x) = \frac{1}{M}\sum_{m=1}^{M} \rho_S(\phi(x),\ \phi(x+\delta_m)) \quad (5)$$

A value of 1.0 indicates a perfectly stable explanation under micro-noise; lower values flag an attribution whose ranking should be cited only with its noise envelope.

TABLE V. FEATURE TAXONOMY USED BY CONSTRAINT_ENFORCEMENT (25 FEATURES PARTITIONED INTO 14 LEGITIMATE + 11 DECISION-LINKED/REPORT-DEPENDENT VARIABLES).

#	Feature	Class	Domain group
1	lens_rise	L	anatomical OCT
2	anterior_depth	L	anatomical OCT
3	aod_nasal	L	anatomical OCT
4	aod_temporal	L	anatomical OCT
5	tisa_nasal	L	anatomical OCT
6	corneal_volume	L	anatomical OCT
7	pupil_x	L	anatomical OCT
8	pupil_y	L	anatomical OCT
9	pupil_diameter	L	anatomical OCT
10	limbus_y	L	anatomical OCT
11	limbus_diameter	L	anatomical OCT
12	age	L	biometric / age
13	k1	L	biometric
14	laterality	L	laterality
15	preop_sph	D	reported refraction
16	selected_sph	D	clinician's selection
17	selected_lens_size	D	clinician's selection (numeric)
18–21	lens_12.1, lens_12.6, lens_13.2, lens_13.7	D	one-hot encoding of the four candidate lens sizes
22–23	ICL, TICL	D	lens-family one-hot
24–25	Myopic, Toric Myopic	D	refractive-type one-hot

3) Operating point: The two sufficiency thresholds are set to $(\tau_{\text{env}}, \tau_{\text{margin}}) = (0.5, 0.3)$. This is the conservative operating point used throughout the study.

The routed-workload aggregate in Section VI-B is also reported at two alternative values of τ_{margin} (Table VI). Routed workload depends on the selected operating point. The architecture defines the aggregate. The deployer chooses the operating point. Among the three operating points reported in Table VI, $\tau_{\text{margin}} = 0.30$ was selected because it routes the fragile mid-band cases that $\tau_{\text{margin}} = 0.20$ leaves unrouted, while holding the human-review workload at 20% (11 of 55), below the 23.6% implied by $\tau_{\text{margin}} = 0.40$; this balances oversight coverage against the review capacity a single clinic can absorb.

B. Output of the Protocol on the Held-Out Cohort

The protocol is run on the 55-eye held-out cohort. Each prediction produces one JSON record with the five core signals, two per-decision oversight fields and case metadata.

The results are read at three levels: the individual decision record, the cohort aggregate used to estimate human-review workload, and the protocol-calibration level where signal distributions are checked on real data. The protocol output is produced at prediction time; if the record is persisted, the same record can later be audited.

Fig. 4 and Fig. 3 reproduce two records as they would appear at decision time, with retrospective outcome information shown separately. Case case4 shows counterfactual evidence supporting deviation from the evaluated lens option. Case case1 shows operational sufficiency despite a later large prediction error, separating decision-time evidence from model-quality assessment.

1) Per-decision evidence: The individual JSON record is the unit at which the deployer documents the output→action relation. Table VII reports five records from the held-out cohort. They were selected because each one shows a different behaviour that the protocol is designed to separate: clean, borderline, flip-prone, out-of-band and attribution-unstable.

Each column of Table VII is one JSON record produced at decision time and persisted for later audit. Two records are reproduced in full in Fig. 3 and Fig. 4. Retrospective measured vault is included only for the discussion of predictive accuracy in Section VII-A.

The five columns show different operational behaviours. The Clean column has every signal above threshold, although its later absolute prediction error is 308 µm. The Borderline and Flip-prone columns show fragile band membership under OCT-grade perturbation, with robustness values of 0.79 and 0.38. The Out-of-band column has stable attribution and band membership, but a negative score margin (-0.38). The Unstable column has the lowest record integrity in the cohort, so its SHAP ranking should be cited only with its noise envelope. A high value on one signal is not used to compensate for failure on another; the axes remain separate.

2) Per-cohort workload: The two cohort-level oversight aggregates in Table VIII convert the 55 decision records into an estimate of human-review workload. This is the cohort-level view relevant to Article 14 oversight. The aggregates are defined as

Definitions. For $N = 55$, the sustainability aggregate is the share of cases with $250 \leq \hat{y}_i \leq 750$. The teaming aggregate is the share of cases satisfying $\text{env}_i \geq \tau_{\text{env}}$ and $\text{margin}_i \geq \tau_{\text{margin}}$. Review required is the complement of the teaming aggregate.

The sustainability aggregate shows that 52 of 55 eyes receive a prediction inside the clinically plausible band. The human-AI-teaming aggregate shows that 44 eyes meet the joint sufficiency criterion and 11 do not. Under the operating point $(\tau_{\text{env}}, \tau_{\text{margin}}) = (0.5, 0.3)$, those 11 eyes represent 20% of the cohort flow. This is the workload routed to human review.

TABLE VI. COHORT ROUTING AGGREGATE UNDER THREE OPERATING POINTS OF THE SCORE-MARGIN SUFFICIENCY THRESHOLD ($\tau_{\text{ENV}} = 0.5$ FIXED).

τ_{margin}	Sufficient (AI-supported workflow)	Routed to review	Rationale
0.20	46 / 55 (0.836)	9 / 55 (0.164)	permissive: only clearly out-of-band cases routed
0.30	**44 / 55 (0.800)**	**11 / 55 (0.200)**	**operating point used in this study**
0.40	42 / 55 (0.764)	13 / 55 (0.236)	conservative: routes mid-band cases for review

Field	Value
CASE	`case1`
Lens option evaluated at decision time	13.2 mm
Predicted vault	498.5 µm
Retrospective measured vault (post-op; not decision-time evidence)	807.0 µm ($\Delta = 308$ µm)
PER-DECISION SIGNALS	
`score_margin`	0.99 (threshold 0.30) ✓ OK
`constraint_enforcement`	0.73 (legitimate share)
`envelope_validity`	1.00 (threshold 0.50) ✓ OK
`decision_robustness`	1.00 (flip rate 0.000)
`record_integrity`	0.86 (rank-order stability)
CONSERVATIVE OPERATION POINT	
`sustainability_impact`	`inside_clinical_band` ✓
`human_ai_teaming`	`ai_supported_workflow_sufficient` ✓
COUNTERFACTUAL BY LENS	
12.1 mm	316 µm margin 0.27 ✓ in band
12.6 mm	351 µm margin 0.40 ✓ in band
13.2 mm (evaluated option)	**499 µm margin 0.99 ✓ in band**
13.7 mm	684 µm margin 0.26 ✓ in band
HUMAN-OVERSIGHT EVIDENCE FUNCTIONS	
14(4)(b) capacities / limits	`score_margin`, `envelope_validity`
14(4)(c) automation bias	`decision_robustness`
14(4)(d) disregard / override / reverse	`constraint_enforcement`, counterfactual
14(4)(e) meaningful intervention	`counterfactual_by_lens`

`predict()` calls only – model internals not accessed

Fig. 3. Decision-time evidence record for case `case1`, with retrospective outcome shown separately. The protocol reports operational sufficiency on all five signals. The measured post-operative vault and absolute error are included only as retrospective outcome information.

Field	Value
CASE	`case4`
Lens option evaluated at decision time	13.7 mm
Predicted vault	845.4 µm
Retrospective measured vault (post-op; not decision-time evidence)	534.0 µm
PER-DECISION SIGNALS	
`score_margin`	-0.38 (threshold 0.30) × FAIL
`constraint_enforcement`	0.46 (legitimate share)
`envelope_validity`	1.00 (threshold 0.50) ✓ OK
`decision_robustness`	1.00 (flip rate 0.000)
`record_integrity`	0.94 (rank-order stability)
CONSERVATIVE OPERATION POINT	
`sustainability_impact`	`outside_clinical_band` ×
`human_ai_teaming`	`review_required` ×
COUNTERFACTUAL BY LENS	
12.1 mm	584 µm margin 0.66 ✓ in band
12.6 mm	531 µm margin 0.88 ✓ in band
13.2 mm	649 µm margin 0.40 ✓ in band
13.7 mm (evaluated option)	**845 µm margin -0.38 × out of band**
HUMAN-OVERSIGHT EVIDENCE FUNCTIONS	
14(4)(b) capacities / limits	`score_margin`, `envelope_validity`
14(4)(c) automation bias	`decision_robustness`
14(4)(d) disregard / override / reverse	`constraint_enforcement`, counterfactual
14(4)(e) meaningful intervention	`counterfactual_by_lens`

`predict()` calls only – model internals not accessed

Fig. 4. Decision-time evidence record for case `case4`, with retrospective outcome shown separately. Status indicators flag the signals failing the operating point of Section VI-A. The measured post-operative vault is included only as retrospective outcome information and is not part of the decision-time record.

TABLE VII. FIVE ILLUSTRATIVE OUTPUT RECORDS FROM THE HELD-OUT COHORT, ONE PER DECISION REGIME. TABLE IS TRANSPOSED FOR READABILITY: EACH COLUMN IS A SINGLE JSON RECORD PRODUCED BY THE PROTOCOL, AND EACH ROW REPORTS ONE FIELD OF THE RECORD. THE COLUMN IS THE AUDIT UNIT: IT IS AVAILABLE TO THE CLINICIAN AT DECISION TIME AND TO THE AUDITOR AFTERWARDS. RETROSPECTIVE MEASURED VAULT IS SHOWN ONLY AS OUTCOME INFORMATION. CASE IDENTIFIERS (`case1`–`case5`) ARE ANONYMISED AND DO NOT CORRESPOND TO CLINICAL-RECORDS DATABASE ENTRIES.

Attribute	Clean	Borderline	Flip-prone	Out-of-band	Unstable
case_id	case1	case2	case3	case4	case5
retrospective vault_real (µm)	807	310	690	534	321
vault_pred (µm)	498	251	774	845	446
score_margin	**0.99**	**0.00**	**-0.09**	**-0.38**	0.78
constraint_enforcement	0.73	0.47	0.79	0.46	0.78
envelope_validity	1.00	1.00	1.00	1.00	1.00
decision_robustness	1.00	0.79	**0.38**	1.00	1.00
record_integrity	0.86	0.93	0.87	0.94	**0.77**
lens option evaluated (mm)	13.2	12.6	13.2	13.7	13.2

TABLE VIII. COHORT-LEVEL OVERSIGHT AGGREGATES AND IMPLIED HUMAN-REVIEW WORKLOAD

Metric	Value	Reading
`sustainability_impact_cohort`	0.945	52 / 55 eyes inside [250, 750] µm
`human_ai_teaming_cohort`	0.800	44 / 55 eyes sufficient under AI-supported workflow
`review_required` (= 1 - teaming)	0.200	11 / 55 eyes routed to human review

Table IX maps each part of the protocol output to a human-oversight or post-market-surveillance function. The table is a traceability map between protocol outputs and the evidence that the deployer must organise and retain.

The clinician uses the record at decision time; persisted records can later be audited and aggregated for post-market surveillance. Both uses rely on `predict()` calls only.

The 11/55 routing result in Table VIII depends on the selected operating point. The deployer calibrates τ_{margin} against two costs: unnecessary routing and unrouted error.

Table VI reports the routed workload at three values of τ_{margin} while keeping τ_{env} fixed at 0.5. The routed workload increases monotonically from 9/55 at the permissive setting to 13/55 at the conservative setting.

3) Signal calibration over the cohort: Table X reports the empirical distribution of the core signals across the held-out cohort. This table gives protocol-level evidence. The clinician and the later auditor use the individual record, as shown in Table VII and Fig. 4 and Fig. 3.

The distribution addresses a methodological objection: the signals could collapse to constants on real data. Here they do not. `score_margin` ranges from -0.38 to 0.99, `constraint_enforcement` from 0.45 to 0.91, `record_integrity` from 0.77 to 0.95, and `decision_robustness` identifies six eyes with flip rates below unity.

4) Envelope-validity calibration: Under the percentile-based threshold, $d_{\max}$ is the 90th percentile of the per-training-point 5-NN mean distance ($d_{\max} = 5.06$). The held-out 5-NN mean distance ranges from 2.20 to 5.97. Fifty eyes remain saturated at `envelope_validity` $= 1.00$; five return values between 0.60 and 1.00. None falls below $\tau_{\text{env}} = 0.5$, so the 11 routed cases are still driven by the score-margin threshold in this cohort. The signal would become more decisive under genuine deployment shift, for example after changes in OCT device, surgeon protocol or patient population.

The cohort mean of `constraint_enforcement` is 0.676. The aggregate attribution share of the legitimate set is 0.681, agreeing to two decimals. Six of the ten highest-ranked features are legitimate, and the largest contributor is the anatomical feature `lens_rise` (Fig. 5). This independent agreement supports the claim that `constraint_enforcement` measures the attribution split it was defined to measure.

C. Counterfactual Over the Discrete Decision Space

Bounded perturbation tests local fragility. It does not compare the four surgical alternatives. The sizing decision is discrete: the surgeon chooses one lens size from four available options. The protocol can therefore query the model under each candidate lens size while holding all other features fixed.

For each held-out eye, the model is queried four times, once per available size. For each query, the protocol recomputes the three sweep-relevant signals: `score_margin`, `envelope_validity` and `decision_robustness`. The result is a 4×55 matrix. Each cell records the predicted vault for one eye under one candidate lens size, together with the three boundary-relevant signals recomputed for that hypothetical decision while all anatomical and biometric inputs remain fixed.

The body of the study reports two slices of the complete 220-row long-format matrix: one decision-level slice and one cohort-level slice.

The decision-level slice gives the operator evidence for disregarding, overriding or intervening under Article 14(4)(d–e). Fig. 4 shows this for case `case4`. The cohort-level slice aggregates across all 55 eyes by lens size and supports post-market surveillance under MDR Article 88. Table XI reports this aggregate.

1) Per-decision counterfactual evidence: Consider eye `case4` at the moment when the surgeon must commit to a lens

TABLE IX. MAPPING OF PROTOCOL OUTPUTS TO HUMAN-OVERSIGHT AND POST-MARKET-SURVEILLANCE FUNCTIONS

Protocol output	Evidence function supported
Per-decision JSON record (Figs. 3, 4)	`Art 14(4)(b)` capacities and limits of the system
`decision_robustness` flag	`Art 14(4)(c)` awareness of automation bias
`constraint_enforcement` taxonomy (legitimate / decision-linked)	`Art 14(4)(d)` basis for disregard, override or reversal
Counterfactual-by-lens matrix (Section VI-C)	`Art 14(4)(e)` basis for meaningful intervention
Cohort routing aggregate (Table VIII)	`Art 14(3)` oversight measures proportionate to risk
Cohort signal distribution (Table X)	`MDR Art 83 & 88` post-market surveillance and trend reporting

TABLE X. DISTRIBUTION OF THE FOUR INFORMATIVE CORE SIGNALS ACROSS THE 55 RECORDS (SIGNAL-CALIBRATION EVIDENCE; NOT PART OF THE PER-DECISION AUDIT UNIT). ENVELOPE_VALIDITY IS CONSTANT AT 1.00. SEE CALIBRATION PARAGRAPH ABOVE.

Statistic	score_margin	constraint_enf	decision_rob	record_int
min	-0.38	0.45	0.38	0.77
Q1	0.42	0.62	1.00	0.81
median	0.65	0.68	1.00	0.85
Q3	0.84	0.74	1.00	0.87
max	0.99	0.91	1.00	0.95
mean $\pm$ std	0.59 $\pm$ 0.33	0.68 $\pm$ 0.11	0.97 $\pm$ 0.10	0.85 $\pm$ 0.05

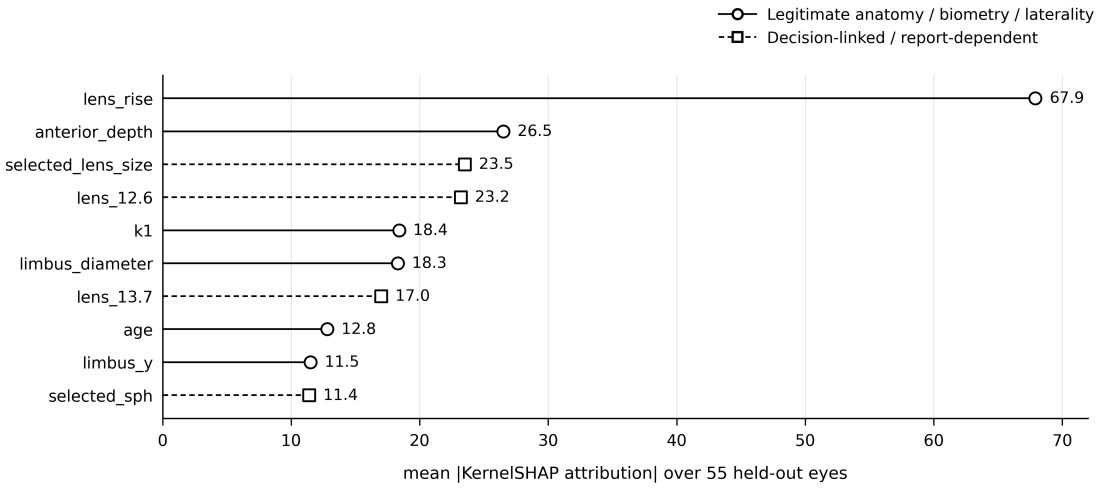


Fig. 5. Top-10 features by mean |KernelSHAP attribution| over the 55-eye held-out cohort. The legitimate/decision-linked split used by `constraint_enforcement` is visible here at the level of individual features.

TABLE XI. COUNTERFACTUAL-BY-LENS-SIZE COHORT AGGREGATES

Lens (mm)	Mean vault (µm)	% in band	Mean margin	Mean robust.
12.1	464.7	98.2	0.66	0.99
12.6	461.6	96.4	0.67	0.98
13.2	542.3	98.2	0.63	0.98
13.7	**684.8**	**63.6**	**0.20**	**0.94**

size. The protocol returns four predictions, one per available size, with their score margin, envelope validity and robustness.

The clinician sees that the 13.7 mm option gives a predicted vault of 845 µm and a score margin of -0.38. The same record also shows that the 12.1 mm and 12.6 mm options produce predictions inside the $[250, 750]$ µm band: 584 µm and 531 µm, with score margins of 0.66 and 0.88.

The decision record and the later audit record are identical. Fig. 4 reproduces the full record. The clinician decides with that evidence available. The auditor later reviews the same evidence.

The counterfactual matrix gives the deployer structured evidence about the available alternatives. It shows the model's prediction under each candidate lens size. This makes intervention under Article 14(4)(e) evidence-based rather than dependent on a single recommendation.

2) Per-cohort post-market surveillance evidence: Aggregating the 4×55 matrix over the held-out cohort reveals a systematic pattern in the deployed regressor (Table XI).

Three of the four lens sizes return predictions inside the clinical band for 96–98% of eyes. The largest size, 13.7 mm, returns predictions inside the band for only 63.6% of eyes. It also has the highest mean predicted vault, 684.8 µm, and the lowest mean robustness in the four-way comparison.

This result should be read as operational monitoring evi-

dence. The deployer does not redesign or retrain the regressor. The value of the matrix is that it identifies a pattern that can be reported through post-market surveillance if it appears in deployment.

The matrix produces this result from `predict()` calls only. It is generated alongside the per-decision telemetry reported in Section VI-B.

Vault standard deviations across the cohort are $\sigma(12.1) = 97.1$ µm, $\sigma(12.6) = 98.1$ µm, $\sigma(13.2) = 109.2$ µm and $\sigma(13.7) = 126.8$ µm.

The counterfactual matrix supports two evidence functions. At decision time, it supports the operator's intervention under Article 14(4)(e). At cohort level, it supports post-market surveillance under MDR Article 88. Both readings are produced from `predict()` calls only, during the time required to predict the four candidate lens sizes, and are persisted with the per-decision telemetry of Section VI-B.

Taken together, the five-signal cohort profile of Section VI-B and the by-lens-size counterfactual matrix of Section VI-C specify the evidence needed by the deployer. The protocol produces per-decision evidence for human oversight under EU AI Act Article 14(3)–(4), and cohort-level signal distributions for post-market surveillance under MDR Articles 83 and 88.

The clinician uses the evidence record at decision time. The auditor and the post-market surveillance process use the same persisted record afterwards. The deployer can produce both readings without access to model internals and without cooperation from the model provider.

D. Reproducibility

The clinical dataset described in Section IV-B cannot be published because it contains personal health information governed by the Medical Devices Regulation and by contractual restrictions across the three contributing centres.

The instrumentation protocol is reproducible at method level. The five core signals are defined operationally in Section VI-A through expressions over `predict()` calls and a fixed background sample. The calibration parameters – clinical band $[250, 750]$ µm, $\sigma = 0.10$, $\varepsilon = 0.01$, $k = 5$, $N = 100$, $M = 50$ and $X_{\text{bg}} = 50$ – are listed in Table IV. The train/held-out split is reproducible with `test_size = 0.10` and `seed = 42`, as stated in Section IV-D.

The deployed model is an Extra Trees regressor over 25 input features partitioned into 14 legitimate and 11 proxy features (Table V); the two sufficiency thresholds are $(\tau_{\text{env}}, \tau_{\text{margin}}) = (0.5, 0.3)$. The protocol is computationally lightweight: feature attribution uses regression-aware KernelSHAP with a 50-sample background, and the perturbation and counterfactual tests require only repeated `predict()` calls, so the full per-decision evidence record is produced without specialised hardware. Experiments were run on Google Colab (CPU runtime) using scikit-learn, NumPy, pandas and SciPy.

A deployer with access to the prediction interface of any regression model on a comparable feature space can rerun the protocol and reproduce the structure of the per-decision evidence reported here. The cohort-level numerical results in Section VI-B and Section VI-C are specific to the contributing cohort; the protocol structure is reusable.

VII. Discussion

A. Operational Sufficiency and Predictive Accuracy are Separate Readings

The main result is that predictive accuracy and operational sufficiency do not collapse into the same judgement.

Case `case3` (Table VII, Flip-prone column) shows the direction in which this separation is most useful. The predicted vault is 774 µm and the absolute prediction error against the measured post-operative vault is 84 µm, which is small by the standards of this cohort. The prediction lies just above the upper edge of the clinical band (`score_margin` = -0.09), but the boundary reading is sharper: `decision_robustness` is 0.38, which means that 62% of bounded perturbations flip the band membership of the prediction. The protocol therefore routes this case to human review even though the regression prediction is accurate by the model's own standards. Good predictive accuracy does not constitute operational sufficiency at the decision boundary.

The reverse direction appears in case `case1` (Fig. 3, Clean column). Every per-decision signal is above its operating threshold, and both per-decision oversight fields report sufficiency. The later measured vault is 807 µm, producing a 308 µm error. The protocol did not fail to predict the outcome; it was not designed to estimate post-operative error after the fact. It was designed to document whether the prediction, at decision time, had the structural properties required for AI-supported workflow.

This separation is methodologically important. Model validation asks whether the regressor predicts vault well enough at population level. Decision-boundary instrumentation asks whether a specific output is supported by enough evidence to be used as a reason for action. Both questions are necessary. They are not interchangeable.

B. Calibration of `Envelope_validity` Under the Percentile Threshold

Under the percentile-based threshold ($d_{\max}$ = 90th percentile of the per-training-point 5-NN mean distance), `envelope_validity` is informative on the present cohort. The threshold value is $d_{\max} = 5.06$ in the standardised feature space. The held-out 5-NN mean distance ranges from 2.20 to 5.97. Fifty of the 55 held-out eyes remain saturated at `envelope_validity` = 1.00; five eyes return values between 0.60 and 1.00. None falls below the sufficiency threshold $\tau_{\text{env}} = 0.5$. The joint criterion of `human_ai_teaming` therefore routes the same 11 cases as the score-margin criterion alone, and the cohort-level workload is unchanged by the percentile calibration.

The informative regime for `envelope_validity` is genuine deployment shift. Changes in OCT device, surgeon protocol or patient population would move more cases beyond the calibrated envelope and yield values below 0.50. Those values would route additional cases through `human_ai_teaming`. A controlled drift experiment on the

input feature space would be the next step for testing that behaviour, but it is outside the scope of this study.

C. One Evidence Record Supports Three Regulatory Evidence Functions

The protocol produces one persisted evidence record that can serve three regulatory evidence functions: decision-time human oversight, later audit and post-market surveillance.

At decision time, the per-decision JSON record (Fig. 4 and Fig. 3) makes the relevant signal values available to the operator. `score_margin` and `envelope_validity` provide evidence for the Article 14(4)(b) function of understanding system capacities and limits in use. `decision_robustness` provides evidence for Article 14(4)(c) by making output fragility visible and reducing blind reliance on the prediction.

The legitimate/decision-linked taxonomy used by `constraint_enforcement` provides evidence for Article 14(4)(d). It shows whether the local prediction is mainly supported by anatomical and biometric variables, or whether it is mainly using variables linked to the clinician's prospective decision. This gives the operator evidence for disregarding, overriding or reversing the output.

The counterfactual-by-lens matrix (Fig. 6, Table XI) provides evidence for Article 14(4)(e). It shows what the model would predict under each available lens size while anatomy and biometry are held fixed. That comparison makes intervention meaningful because the operator can inspect the available alternatives, not only the evaluated option.

After the decision, the same persisted record becomes an audit unit. Across cases, the same records become post-market surveillance evidence under MDR Articles 83 and 88. The 13.7 mm pattern in Table XI is a concrete example: it is not a single adverse event, but a cohort-level signal that could justify closer monitoring if reproduced in deployment.

D. Boundary Conditions for Deployer-Side Use

The protocol should be read under four boundary conditions. First, decision-boundary instrumentation presupposes a model that has already passed model-level validation. The signals reported here evaluate whether a specific output has the observable properties required for use in a workflow. Population-level validation, clinical evaluation and model selection remain prerequisites. In this case, that prerequisite is represented by the regression comparison and cross-validated MAE reported in Table II.

Second, several elements of the protocol require explicit configuration before deployment. The legitimate/decision-linked feature taxonomy defines how `constraint_enforcement` is interpreted. The KernelSHAP background sample defines the reference distribution for local attribution. The operating thresholds define the review workload accepted by the organisation. They are explicit elements of the deployer-side evidence package and should be documented before operational use.

Third, the empirical results are case-study results from a 55-eye held-out cohort within one multi-site ophthalmology network. The workload estimate, the 13.7 mm counterfactual pattern and the observed signal distributions should therefore be read as evidence of how the protocol behaves in this deployment setting. Transfer to other ICL datasets, devices, surgeons or patient populations would require recalibration and prospective monitoring.

Fourth, the protocol specifies five core signals derived from four documented failure modes; it does not claim that these five are necessary and sufficient. Failure modes that the AI-safety literature has identified but that the present protocol does not instrument – for instance, calibration drift, label distribution shift or systematic disagreement with an expert reference – would require additional signals not covered here. The heuristic parameter values reported in Box 2 reflect operational choices grounded in the cited literature; a systematic sensitivity study across all heuristic parameters, beyond the τ_{margin} axis reported in Table VI, is left for future work.

VIII. Conclusion

This study treats post-hoc XAI as deployer-side evidence rather than as explanation alone. Much of the XAI literature characterises the model, which is mainly the developer's question. Regulated deployment creates a different question for the deployer: when a prediction supports an action, what evidence is attached to that prediction, and can another party audit it later?

The study specifies a decision-boundary instrumentation architecture that answers that question by reorganising existing XAI and validation methods – KernelSHAP, k-nearest-neighbour distance, bounded perturbation and rank-order stability – into a per-decision telemetry record. The record is computed from `predict()` calls only and is designed to support decision-time human oversight, later audit and post-market surveillance.

The protocol was applied to a 55-eye held-out cohort of an ICL vault prediction regressor. At the conservative operating point, 11 cases (20%) were routed to human review and the rest were classified as operationally sufficient. Each case produced a JSON record that maps to deployer obligations under EU AI Act Article 14 and to MDR post-market surveillance reporting under Articles 83 and 88. The counterfactual-by-lens matrix showed how the same prediction interface can support meaningful intervention at decision time and cohort-level monitoring afterwards.

These results hold under clear limitations. The empirical findings come from a single 55-eye held-out cohort in one ophthalmology network, so the workload estimate and counterfactual pattern are cohort-specific and require multi-site, prospective validation before generalisation. Some signals saturate in this cohort: `envelope_validity` equals 1.00 for 50 of the 55 cases (Section VII-B), so its discriminative contribution should be re-assessed under more varied deployment conditions. The protocol is also not yet compared against alternative evidence-generation approaches such as confidence-threshold routing or clinical audit checklists. Consolidating the future work noted above, the main directions are: multi-site and prospective validation, a systematic sensitivity study beyond the τ_{margin} axis, comparison against baseline evidence-generation approaches, and additional signals for failure modes

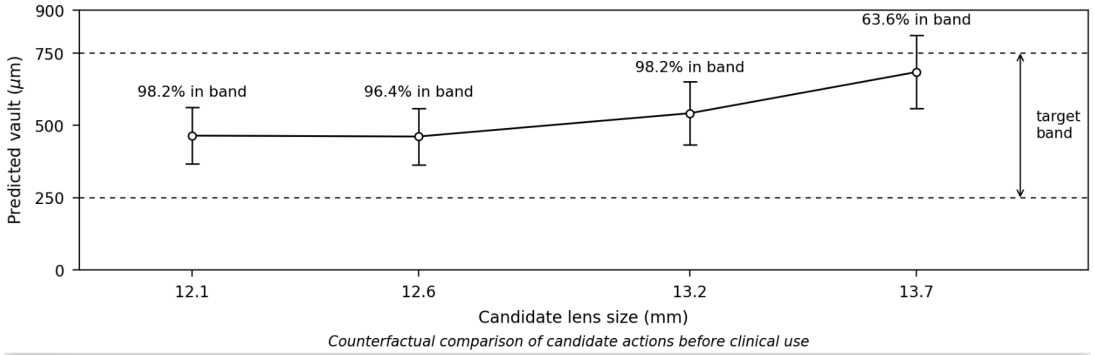


Fig. 6. Cohort-level visualisation of the by-lens-size counterfactual matrix of Table XI. Each bar is the mean predicted vault under one candidate lens size, computed across all 55 held-out eyes; the vertical lines mark ± one standard deviation. The horizontal shaded band marks the clinical sufficiency interval [250, 750] µm. The three smaller lens sizes place their mean predictions near the centre of the band; the 13.7 mm option pushes the mean towards the upper edge, with the largest standard deviation in the four-way comparison and the lowest in-band fraction (63.6%). The pattern is generated from `predict()` calls only and is the visual companion to the post-market surveillance reading of Section VI-C.

not instrumented here (for example calibration drift, label-distribution shift, or systematic disagreement with an expert reference).

The main methodological claim is limited. Under limited-inspection conditions, XAI can function as instrumentation without requiring access to model internals: it characterises observable behaviour, tests boundary stability and creates technical trust evidence at the moment when an output may become a reason for action.

ACKNOWLEDGMENT

This research has been supported by the Ministerio de Ciencia, Innovación y Universidades, Gobierno de España, under research grant PID2024-158490OB-C33 (ECEAAS).

DECLARATION ON GENERATIVE AI

The authors used Claude (Anthropic, `claude-opus-4`) to assist in drafting prose from author-specified outlines and in generating Python code and replication scripts. The development of the scientific ideas, theoretical framework, research design, results and conclusions was carried out by the authors. The authors critically reviewed and edited all AI-assisted text and code, and assume full responsibility for the accuracy, originality and integrity of the manuscript. No generative AI tool is listed as an author.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence," Official Journal of the European Union, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] Co of the European Union and European Parli, "Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices," Official Journal of the European Union, 2017. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2017/745/oj

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.

[4] M. Ouchi *et al.*, "Vault changes in eyes with a vertically implanted implantable collamer lens," *Scientific Reports*, 2024, clinical reference using the commonly cited ICL vault safety interval of approximately 250–750 micrometres.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[6] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[7] A. B. Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 9505–9515.

[9] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[11] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[12] Q. V. Liao and K. R. Varshney, "Human-centered explainable AI (XAI): From algorithms to user experiences," arXiv:2110.10790, 2022.

[13] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[14] Artificial Intelligence Board and Medical Device Coordination Group, "Aib 2025-1 / mdcg 2025-6: Interplay between the medical devices regulation (mdr) & in vitro diagnostic medical devices regulation (ivdr) and the artificial intelligence act (aia)," European Commission guidance document, Jun. 2025. [Online]. Available: https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en

[15] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, 2016.

[16] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

[17] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable algorithms," *University of Pennsylvania Law Review*, vol. 165, no. 3, pp. 633–705, 2017.

[18] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, 2023.

[19] International Organization for Standardization, "ISO/IEC 42001:2023: Information technology — artificial intelligence — management system," International standard, 2023. [Online]. Available: https://www.iso.org/standard/81230.html

[20] B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen, "Beware explanations from AI in health care," *Science*, vol. 373, no. 6552, pp. 284–286, 2021.

[21] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proceedings of the 4th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 106, 2019, pp. 359–380.

[22] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 310, 2020.

[23] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

[24] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6:1–6:13, 2019.

[25] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[26] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33–44.

[27] J. Singh, J. Cobbe, and C. Norval, "Decision provenance: Harnessing data flow for accountable systems," *IEEE Access*, vol. 7, pp. 6562–6574, 2019.

[28] D. Urrutia-Onate, A. Perallos, and E. Onieva, "XAI as decision-boundary instrumentation for AI governance," Zenodo preprint, 2026.

[29] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, 2011.

[30] J. Rasmussen, "Risk management in a dynamic society: A modelling problem," *Safety Science*, vol. 27, no. 2–3, pp. 183–213, 1997.

[31] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010.

[32] Y. Shen, L. Wang, W. Jian *et al.*, "Big-data and artificial-intelligence-assisted vault prediction and EVO-ICL size selection for myopia correction," *British Journal of Ophthalmology*, vol. 107, no. 2, pp. 201–206, 2023.

[33] A. Russo, O. Filini, G. Savini, G. Festa, F. Morescalchi, A. Boldini, and F. Semeraro, "Predictability of the vault after implantable collamer lens implantation using OCT and artificial intelligence in white patient eyes," *Journal of Cataract & Refractive Surgery*, vol. 49, no. 7, pp. 724–731, 2023.

[34] P. Zeboulon *et al.*, "Validation of a new implantable collamer lens sizing algorithm based on SS-OCT images," *Journal of Cataract & Refractive Surgery*, vol. 52, no. 1, pp. 61–66, 2026.

[35] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] H. Yuan, M. Liu, L. Kang, C. Miao, and Y. Wu, "An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models," 2022.

[39] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 162, 2022, pp. 20 827–20 840. [Online]. Available: https://proceedings.mlr.press/v162/sun22d.html

DcDM: A Pre-training Data Evaluation Framework for Proactive Drift Prevention in Machine Learning

Okjoo Choi[1], Wonsun Shin[2]
Dept. of AI·Software Engineering, PAI CHAI University, Daejeon, Korea, Republic of Korea[1]
Dept. of Smart ICT Convergence, PAI CHAI University, Daejeon, Korea, Republic of Korea[2]

***Abstract*—The performance of machine learning (ML) systems often deteriorates over time owing to data drift, which is typically caused by changes in data quality or distribution. Such degradation in deployment environments can result in inaccurate predictions and reduced system reliability. Conventional drift detection approaches have largely focused on retraining ML models after performance degradation has occurred. However, because the root causes of drift often originate from the data itself, a data-centric approach is needed to address the problem at its source. This study proposes a systematic, data-centric drift management (DcDM) framework that integrates domain-specific rule validation, data quality assessment, and statistical drift analysis before model training. By first verifying semantic constraints and then assessing data quality and distributional stability, DcDM enables early identification of potential drift hazards and prevents low-quality or semantically invalid data from entering the training pipeline. We evaluate the proposed framework on three datasets across different modalities: Electricity Load Diagrams, CIFAR-10, and VisDA-2017. The experimental results show consistent improvements of approximately 6% to 12.5% in both accuracy and F1-score. Additional ablation studies, baseline comparisons, and statistical significance tests further demonstrate the robustness and effectiveness of the proposed approach.**

***Keywords*—Data drift; data evaluation; data quality; domain rule; drift prevention and mitigation; drift management**

I. Introduction

Machine learning (ML) systems have become indispensable across a wide range of applications, including image classification, speech recognition, financial forecasting, and medical diagnostics. As ML adoption expands across diverse industrial domains, ensuring the accuracy and long-term reliability of deployed models has become increasingly critical [1]. A primary challenge faced by these systems is data drift, a phenomenon in which the statistical properties of input data change over time, potentially degrading model performance [2]. For example, a model trained to predict a company's stock price using historically stable market data may initially perform well; however, as market volatility increases, the shifting data distribution can render inaccurate predictions [3]. Such drift-induced degradation is especially problematic in real-world environments where continuous reliability is paramount [4].

Traditional responses to data drift typically adopt model-centric strategies, such as retraining or fine-tuning models once performance decline is observed [5]. While these methods can offer temporary recovery, they are frequently time-consuming and computationally expensive, particularly in dynamic environments where operational data changes rapidly [6]. Moreover, such approaches fail to address the root cause of drift, which frequently originates from the data itself rather than from the model [7]. In contrast, data-centric drift management emphasizes maintaining data quality and semantic integrity throughout the ML lifecycle [8][9]. This paradigm focuses on the systematic evaluation of data properties to proactively detect and mitigate drift before it impacts model performance [10].

In this study, we propose DcDM, a unified pre-training framework for proactive drift prevention. The proposed framework is not the introduction of isolated techniques for drift management, but the integration of domain rule validation, data quality assessment, and statistical drift analysis into a single upstream pipeline. This design differs from conventional MLOps and drift-monitoring tools, which are primarily reactive and operate after deployment. DcDM is intended to detect drift hazards before model training and to filter or refine problematic data before they affect downstream learning. This study makes four contributions. First, it introduces a unified data-centric framework that combines domain rule validation, quality assessment, and drift analysis in a single workflow. Second, it operationalizes this framework as an end-to-end pipeline for pre-training data evaluation. Third, it demonstrates the framework on heterogeneous datasets spanning tabular time-series and image domains. Fourth, it performs controlled ablation and baseline comparisons to quantify the contribution of each pipeline component and to distinguish DcDM from simple data cleansing.

To demonstrate the effectiveness and generalizability of the proposed framework, we conducted experiments on three benchmark datasets from different domains: electricity load diagrams (time-series), CIFAR-10 (image classification), and VisDA-2017 (domain adaptation). The experimental results reveal significant improvements, ranging from approximately 6% to 12.5% in both accuracy and F1-score, across all datasets. These findings highlight the utility of proactive data evaluation in preserving the predictive performance and mitigating the risk of drift-induced degradation in ML systems.

II. Related Work

Managing the performance degradation caused by data drift has been extensively studied in recent years. The existing research can be broadly categorized into three areas: 1) drift types, 2) drift detection methods, and 3) drift management frameworks, including data-centric approaches.

A. Drift Types in ML

Drift in ML systems refers to the change in data properties over time, resulting in a potential mismatch between training and operational environments. There are two primary types of drift: data drift (covariate shift) and concept drift. Data drift involves a change in the distribution of input features *P(x)*, whereas concept drift reflects changes in the conditional distribution *P(y|x)* over time.

1) Data drift: Data drift is defined as a change in the distribution of input data relative to the distribution present during model training.

$$P_{tran}(x) \neq P_{operational}(x)$$

For example, noise due to age-related deterioration of camera lenses or out-of-focus images constitutes data drift, which affects classification performance.

2) Concept drift: Concept drift occurs when the relationship between input and target variables changes relative to the relationship learned during model training.

$$P_{tran}(y|x) \neq P_{operational}(y|x)$$

For example, a machine learning model trained to detect spam emails based on content may fail when spam patterns evolve significantly. Concept drift can be further categorized into four main categories: sudden, gradual, incremental, and recurring drift [11].

B. Drift Detection Methods

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

Early drift detection studies mainly focused on statistical monitoring approaches that compare operational data distributions with baseline training distributions.

1) Statistics methods: Statistical test methods measure distribution changes between two datasets (training vs. current). These include:

a) Population Stability Index (PSI): PSI assesses whether training and operational data come from the same distribution [12]. Lower PSI values indicate more similar, stable distributions.

b) Kullback-Leibler(KL) Divergence: KL divergence quantitatively measures the extent to which the distribution of operation data differs from training data [13]. An increase beyond a threshold signals drift.

c) Jensen-Shannon(JS) Divergence: JS divergence is a symmetrized and smoothed version of KL divergence, bounded between 0 and 1, providing a normalized similarity measure [14].

d) Kolmogorov-Smirnov(KS) Test: The KS test compares cumulative distribution functions (CDFs) of two datasets. A significant p-value (< 0.05) indicates distributional differences signalling potential data drift [15].

2) Streaming methods: Streaming methods monitor statistical measures such as mean, variance, skewness, and kurtosis over time. These include:

a) Drift Detection Method (DDM): [16]

b) Early Drift Detection Method (EDDM) [17]

c) Exponentially Weighted Moving Average (EWMA) [18].

3) Time window-based approaches: These methods typically use a fixed reference window, summarize historical information, and a sliding window that summarizes up-to-date information. Dividing data into time windows and comparing statistics or distributions across these windows can help detect gradual changes in the data over time. A significant difference in the distributions of these windows indicates that a drift has occurred. These methods include adaptive windowing (ADWIN) [19] and hopping drift detection (HDDMA and HDDMW) tests [20].

4) Representation-space drift detection:

a) MMD(Maximum Mean Discrepancy): MMD is a kernel-based statistical test that measures the distance between the mean embeddings of two distributions in a reproducing kernel Hilbert space. It is particularly suited to high-dimensional data such as images and text, and has been applied to drift detection in deep learning contexts [21].

b) FID(Fréchet Inception Distance): FID measures the distance between feature distributions of real and generated (or reference vs. operational) image datasets by comparing the mean and covariance of Inception-v3 feature vectors. It is widely used in image-based drift detection [22].

These drift detection methods are widely used for structured data, whereas unstructured data often requires perceptual or embedding-based approaches. Table I presents the classification of drift detectors by data type.

TABLE I. DRIFT DETECTION METHOD BY DATA TYPES

Data Type	***Method***	***Reason***
Structured (Tabular)	PSI, KS-Test, KL/JS, ADWIN, DDM, EDDM	Well-suited for numerical and categorical data with a defined schema
Image	FID, LPIPS, MMD, KL/JS, HDDDM, Classifier-based	Support perceptual similarity and high-dimensional shifts
Audio	KL/JS, MMD, Page-Hinkley, Classifier-based	Use spectrograms or embeddings for detection
Text	KL/JS (embeddings), MMD, Classifier-based	Apply embedding models like BERT for drift detection
Streaming	ADWIN, DDM, EDDM, Page-Hinkley	Real-time monitoring of concept drift

C. AutoML-Based Method

Celik proposed strategies to address conceptual drift in AutoML environments using Bayesian optimization and genetic programming [23]. AML4S is a framework that automates ML model selection and hyperparameter tuning to adapt the model in real time when performance drops below a threshold (approximately 7%) [24]. An AutoML-DC model utilizing

meta-learning and automated ensemble learning was proposed to correct for incremental conceptual drift in sensor data [25]. These studies remain focused on model retraining and address issues within limited scopes, such as IoT environments or sensor data.

D. Data-Centric AI and MLOps

The data-centric AI paradigm, advocated by Ng [26], shifts focus from model architecture optimization to systematic improvement of data quality, labeling, and validation. In this view, fixing the data yields more reliable gains than iterating on the model alone. MLOps frameworks such as Evidently AI [27] and Alibi-Detect [28] provide open-source tooling for monitoring data distributions and detecting drift in production pipelines. However, these tools primarily operate reactively—alerting practitioners after drift symptoms emerge—and do not integrate domain-specific semantic rule validation into a unified pre-training pipeline. DcDM complements and extends these tools by adding upstream semantic validation and quality gating before model training.

III. Research Background

Although considerable progress has been made in drift detection and MLOps automation, existing solutions often address only isolated stages of the machine learning lifecycle. Most prior methods focus on model retraining or statistical detection after drift symptoms have already manifested in deployed systems. Consequently, these approaches tend to be reactive rather than proactive, which can lead to delayed corrective actions and inefficient model maintenance cycles.

In addition, conventional drift-handling frameworks seldom incorporate domain-specific constraints or data quality metrics at the pre-modeling stage. This lack of early validation often results in the propagation of low-quality or semantically invalid data into the training pipeline, thereby compromising the robustness and generalization ability of the model.

To overcome these limitations, we introduce a data-centric drift management (DcDM) framework that proactively monitors and evaluates data throughout the ML lifecycle. The DcDM framework aims to:

- Prevent drift hazards by validating data against business and domain-specific rules (semantic validation).
- Quantify drift risks by assessing quality metrics and statistical indicators before model training.
- Enable early interventions, such as data cleansing or discarding, before drift-induced performance degradation occurs.

This approach is grounded in the data-centric AI paradigm, which prioritizes data quality and integrity over model-centric optimization alone. Our contribution is the integration of rule-based evaluation, quality assessment, and statistical drift detection into a unified workflow, enabling ML systems to be more resilient to dynamic, real-world data environments. Table II lists the key terms defined in this study.

TABLE II. Key Terms

Term	Definition
Drift Hazard	A potential anomaly in the input data that could lead to degraded ML performance, identifiable before model training or inference.
Drift Risk	The probability that a drift hazard will actually lead to performance degradation during or after prediction.
Drift Prevention	A proactive strategy to detect and assess incoming data for potential hazards. If anomalies are detected—whether due to shifts in data distribution or quality violations—the system responds by collecting new data or retraining the model to realign with the original objectives.
Drift Mitigation	The process of excluding data that fails domain rule validation or quality checks from model training and inference, thereby minimizing the risk of drift.

Fig. 1 illustrates potential risks and hazards associated with drift across the ML lifecycle.

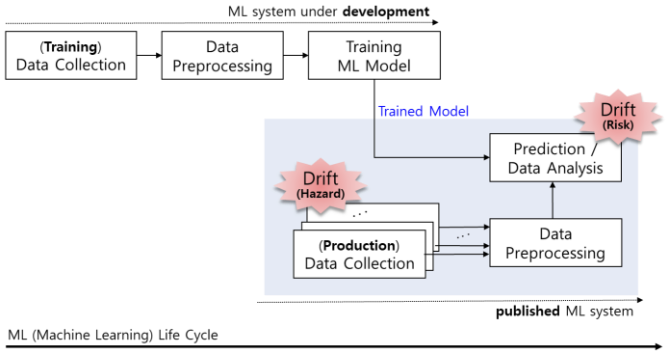


Fig. 1. Drift hazard and risk.

In contrast to previous studies that focus on detecting and responding to drift after performance deterioration, our approach emphasizes early-stage intervention. By systematically assessing data before it influences the model, our framework enables both drift prevention and effective mitigation.

IV. Data-Centric Drift Management

This study proposes a data-centric drift management (DcDM) framework that integrates data validation and drift detection into the ML pipeline. The proposed framework is structured into three steps—data requirement, data evaluation, and model performance analysis—to ensure that only clean, reliable data proceed to model training and inference. The data evaluation process comprises three key components: domain rule validation, data quality assessment, and drift analysis. This stepwise data evaluation aims at ensuring semantic validity by enforcing domain-specific constraints and confirming technical reliability using established quality metrics before drift detection. The DcDM framework is shown in Fig. 2, effectively prevents drift hazards and identifies and mitigates the performance degradation caused by drift.

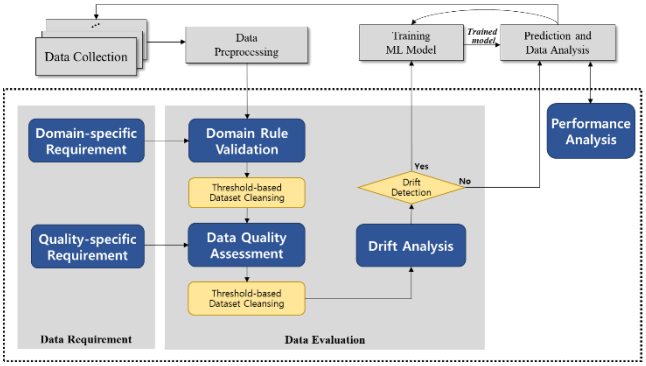


Fig. 2. Data-centric Drift Management (DcDM) framework.

A. Data Requirement

The first step establishes explicit data requirements that incorporate both business-specific rules and quality-based constraints. These requirements serve as a foundation for subsequent validation, monitoring, and drift mitigation activities.

1) Domain-specific requirement: Domain-specific requirements define the business-driven constraints that the dataset must satisfy. These are formalized as domain rules DR_i, which define valid value ranges [α_low_i, α_high_i], or logical inter-attribute relationships. Rules are typically derived from domain expertise, empirical guidelines, or regulatory standards and serve as primary checkpoints for semantic validity. Beyond univariate constraints, inter-attribute rules (e.g., the number of confirmed infection cases must not exceed the number of individuals tested) preserve relational integrity.

2) Quality-specific requirement: High-quality data is essential for preventing data drift. This study adopts goal-based data quality metrics (DQM) tailored to each application domain. For structured data as Table III, four metrics are quantified: accuracy, completeness, consistency, and redundancy.

TABLE III. STRUCTURED DATA: QUALITY METRICS AND THRESHOLDS

Metrics	Threshold	Violation Rate and Judgment		
		Keep (≤)	***Cleansing*** (≤)	***Discard*** (>)
Accuracy	0.95	5%	15%	15%
Completeness	0.95	5%	20%	20%
Consistency	0.95	10%	30%	30%
Redundancy	0.95	5%	20%	20%

For unstructured data such as images, audio, and text, domain-adapted metrics assess semantic integrity, content usability, and statistical coherence, as Table IV. Note that for low-resolution tasks (e.g., CIFAR-10 at 32×32), resolution thresholds and perceptual metrics are adjusted accordingly to reflect realistic quality expectations for the target data modality.

B. Data Evaluation

In this step, data are assessed to ensure compliance with the goal-oriented requirements defined by domain-specific rules and quality criteria. Subsequently, drift detection is performed on data meeting these criteria. The key idea is to pre-evaluate the integrity and usability of the operational dataset (D_op) using the clean reference dataset (D_ref) as a benchmark before training or inference.

TABLE IV. UNSTRUCTURED DATA: QUALITY METRICS AND THRESHOLDS

Data Type	Metric	Description	Threshold / Evaluation
Text	Readability Score	Assesses sentence clarity and grammar	≥ 60 readability score
	Language Consistency	Consistent language use across documents	≥ 95% consistent tokens
Image	Resolution Quality	Pixel density meets the application requirement	≥ 224x224 pixels
	Clarity / Blur Score	Measured using SSIM or Laplacian variance	SSIM ≥ 0.90 or LapVar ≥ 100
	Redundancy Detection	Duplicate image detection	≤ 5% duplicates
Audio	Signal-to-Noise Ratio	Quantifies audio clarity vs noise	SNR ≥ 20 dB
	Duration Validity	Consistent duration across samples	[1s – 10s]
	Transcription Accuracy	ASR transcription confidence	≥ 95% WER-based confidence

Algorithm 1 formalizes the three-step evaluation process. Thresholds α, β, and γ are selected empirically using a validation holdout set: the optimal values (α=10%, β=10%, γ=0.05) were identified through a grid search over the Electricity dataset and subsequently confirmed via the sensitivity analysis reported in Section V-E.

Algorithm 1: Data Evaluation

Input
  D_ref: reference dataset (baseline)
  D_op: operational dataset (current or incoming)
  DR: set of domain rules {DR_1, DR_2, ..., DR_n}
  DQM: set of data quality metrics {DQM_1, DQM_2, ..., DQM_m}
  α: rule violation threshold
  α_high: rule violation discard threshold
  β: data quality threshold
  β_high: data quality discard threshold
  γ: drift detection threshold
Output:
  Decision on D_op: {Accept, Cleansing, Discard}
  Drift Report: {Type, Magnitude, Location}

Step I: Domain Rule Validation
  for each domain rule DR_i in DR do
    ViolationRate[DR_i] ← EvaluateViolation(D_op, DR_i)
  end for
  v_total ← Average(ViolationRate)

  if v_total > α then
    if v_total > α_high then
      return Discard, Reason: Critical domain rule violations"
    else
      D_op ← CleansingData(D_op, DR)
    end if
  end if

Step II: Data Quality Assessment

```
    for each quality metric DQM_j in DQM do
      score[DQM_j] ← EvaluateQuality(D_op, DQM_j)
      Δ[DQM_j] ← CompareWithReference(D_ref, score[DQM_j])
    end for
    Δ_total ← Aggregate(Δ)

    if Δ_total > β then
      if Δ_total > β_high then
        return Discard, Reason: Severe quality degradation
      else
        D_op ← CleansingData(D_op, DQM)
      end if
    end if

Step III: Drift Analysis
    for each feature f of dataset do
      DriftResult.Score[f] ← Distance(D_ref(f) , D_op(f))
      if DriftResult.Score[f] > γ then
        ReportDrift(DriftResult)
        return Discard, Reason: Drift Detected, Go to Model
Retraining
      endif
    end for
    return Accept, Reason: No significant drift, Go to Prediction
```

1) Domain rule validation: This component applies predefined domain-specific rules derived from expert knowledge. The violation rate for each rule DR*i* is computed as:

$$\text{V_total}() = \frac{number\ of\ records\ violating\ DR_i}{total\ number\ of\ records} \quad (1)$$

The average violation rate V_total is compared against the threshold α. Based on V_total, datasets are classified as accept (< 5% violation), cleansing (5%–15%), or discard (> 15%). Unlike traditional drift detection that analyses whole-dataset distributions, domain rule verification focuses on whether individual key attributes adhere to predefined constraints, enabling early, fine-grained detection of data anomalies.

2) Data quality assessment: A suite of quality metrics (DQM) is computed on D_op. The deviation from D_ref for each metric DQM_j is:

$$\Delta[DQM_j] = |\, DQM_j(D_op) - DQM_j(D_ref) \,| \quad (2)$$

where, Δ is the individual quality deterioration per metric.

These deviations are aggregated (via mean or weighted sum) to obtain a cumulative degradation score, Δ_total. Δ_total ← Aggregate(Δ) summarizes the overall quality degradation across all metrics as shown in Eq. (3):

$$\Delta\text{_total} = \left(\frac{1}{m}\right) \sum \Delta[DQM_j] \quad \text{for } j = 1 \text{ to } m \quad (2)$$

where, *m* is the number of quality metrics.

If Δ_total exceeds a threshold (β), the dataset is refined or discarded. A sudden drop in completeness or spike in redundancy signals anomalies during data acquisition or preprocessing. Datasets classified as Discard at this stage are excluded from training; practitioners must collect or re-generate a compliant dataset before proceeding.

3) Drift analysis: Distributional shifts are monitored using statistical drift detection methods appropriate to each data modality (KS test for tabular data; FID and cosine similarity for images). If the drift score for any feature *f* exceeds threshold γ, the framework triggers model retraining. The per-feature drift score uses the same notation throughout: DriftScore[f] denotes the distributional distance for feature *f* between D_ref and D_op.

C. Model Performance Analysis

Only datasets satisfying all three evaluation criteria proceed to model training or inference. Model performance is continuously monitored; a significant decrease in accuracy or F1-score may indicate a change in the underlying data generation process, inducing model retraining. By conducting data evaluation at the early stages of the ML lifecycle, DcDM reduces the frequency of costly real-time model retraining.

V. Experiments

To validate the effectiveness of the DcDM framework, we conducted experiments using three datasets from different modalities: Electricity Load Diagrams (tabular time-series), CIFAR-10 (image), and VisDA-2017 (image for domain adaptation). For each dataset, multiple operational variants were generated to simulate different types of drift. Each dataset undergoes domain rule evaluation, data quality evaluation, and drift detection sequentially. We then compare model performance before and after refinement, conduct an ablation study to isolate each component's contribution, and compare against established drift detection baselines. All improvements are confirmed to be statistically significant.

A. Experimental Setup

1) Datasets

a) UCI Electricity load diagrams: Time-series dataset capturing electricity consumption by households, used to evaluate time-based drift patterns [29].

b) CIFAR-10: A standard image classification benchmark containing 10 classes of low-resolution (32×32) color images [30].

c) VisDA-2017: A domain adaptation benchmark with synthetic-to-real domain shift in object recognition tasks [31].

2) Experiment Scenario

A systematic experiment scenario was designed as follows:

a) Operational dataset generation: For each dataset, multiple operational variants (e.g., op1, op2) were synthetically generated by applying random transformations to simulate real-world anomalies as shown in Table V. These perturbations were selected to approximate common operational anomalies in each modality while still allowing controlled comparison across datasets.

b) Domain rule evaluation: A customized set of domain-specific rules and threshold values was defined per dataset. Datasets with violation rates below 10% were retained; those between 10%–20% were cleaned; those exceeding 20% were discarded.

TABLE V. SYNTHETIC PERTURBATION AND PARAMETER

Modality	Perturbation	Parameterization (sampled per variant)
Tabular (Electricity)	Missing-value injection	Uniform random masking of 2%–25% of cells, applied independently per column
	Power-surge injection	Additive spikes of 1.5×–3×the column's local rolling mean, inserted at 1%–10% of timestamps
	Usage-pattern shift	Multiplicative rescaling of weekend or off-peak readings by a factor drawn from U(0.5, 2.0)
Image (CIFAR-10 / VisDA-2017)	Brightness distortion	Pixel-wise additive/multiplicative shift, target mean brightness drawn from U(40, 220) on a 0–255 scale
	Gaussian noise injection	σ drawn from U(5, 35) on a 0–255 pixel scale, applied per-channel
	Blurring	Gaussian blur kernel with σ drawn from U(0.5, 3.0)
	Label/channel corruption	Label index reassignment for a sampled fraction (0%–30%) of records; for VisDA-2017, additional RGB channel permutation at a sampled probability

c) Data quality assessment: Quality metrics and thresholds tailored to each dataset type assessed completeness, consistency, and noise levels.

d) Drift analysis: PSI, JS Divergence, KL Divergence, and KS test were applied to tabular datasets; SSIM, PSNR, NIQE, BRISQUE, and FID were used for image datasets.

e) Model performance comparison: Model performance on each refined dataset was compared against its unrefined counterpart.

f) Ablation study: Each pipeline stage (domain rule validation, quality assessment, drift detection) was removed individually to quantify each component's contribution.

g) Baseline comparison: DcDM was compared against ADWIN, DDM, and Evidently AI on the Electricity and CIFAR-10 datasets to contextualize performance gains.

B. Case I: Electricity Load Diagrams

1) Operational dataset generation: Ten operational datasets (op1_elec to op10_elec) were generated from the electricity dataset, simulating anomalies including missing values, power surges, and unusual usage patterns.

2) Domain rule evaluation: Domain rules and thresholds for the Electricity dataset are defined in Table VI.

TABLE VI. DOMAIN RULES AND THRESHOLDS (ELECTRICITY)

NO	Rule	Threshold	Description
DR1	Power ≥ 0	≥ 5%	Power values must be non-negative
DR2	0 ≤ Valid Power ≤ 5000kW	≥ 10%	Valid power values within the operating range
DR3	Weekend power < weekday avg.	< 500	Weekend power usage must be lower than the weekday average
DR4	Hourly change rate	≥ 15%	Hourly change rate must be within ±20%

The domain rule evaluation results are presented in Table VII. Datasets receiving a 'Cleansing' judgment had violating records removed before proceeding to quality assessment.

TABLE VII. DOMAIN RULE EVALUATION (ELECTRICITY)

Dataset	DR Violations	Judgment	New Dataset
op1_elec	DR1: 2%, DR4: 18%	Cleansing	op1-1_elec
op2_elec	DR2: 12%	**Discard**	-
op3_elec	DR3 Violation	Cleansing	op3-1_elec
op4_elec	DR4: 25%	Cleansing	op4-1_elec
op5_elec	DR2: 5%, DR4: 10%	Keep	
op6_elec	DR1: 7%	**Discard**	-
op7_elec	DR4: 16%	Cleansing	op7-1_elec
op8_elec	DR3 Violation, DR4: 22%	Cleansing	op8-1_elec
op9_elec	DR2: 18%	**Discard**	-
op10_elec	DR4: 14%	Keep	

3) Data quality assessment: Quality thresholds for the Electricity dataset: Completeness ≥ 95%, Consistency ≥ 98% (no duplicate timestamps), Timeliness ≥ 97% (records within expected intervals). Results are presented in Table VIII.

TABLE VIII. DATA QUALITY ASSESSMENT (ELECTRICITY)

Dataset	Comp	Cons	Time	Judgment	New Dataset
op1-1_elec	94%	99%	98%	Cleansing	op1-2_elec
op3-1_elec	96%	98%	99%	Keep	
op4-1_elec	98%	97%	97%	Keep	
op5_elec	99%	98%	96%	Keep	op5_elec
op7-1_elec	95%	96%	94%	Cleansing	op7-2_elec
op8-1_elec	93%	94%	97%	Cleansing	op8-2_elec
op10_elec	99%	99%	98%	Keep	

4) Drift analysis: The KS test was used to detect drift between refined operational datasets and the original data. As shown in Fig. 3, a p-value below 0.05 indicates drift. Most unrefined datasets (op1_elec, op3_elec, op4_elec, op7_elec, op8_elec) showed significant drift, while their refined counterparts showed substantially lower drift probability. The op5_elec dataset exhibited borderline drift (p-value = 0.039) despite passing domain rule and quality checks, triggering model retraining.

5) Model performance comparison: Using a Random Forest classifier, average model accuracy improved from 0.821 to 0.869, and F1-score increased from 0.824 to 0.872, representing relative improvements of 5.85% and 5.83%, respectively. Fig. 4 shows the model performance on the Electricity operational datasets.

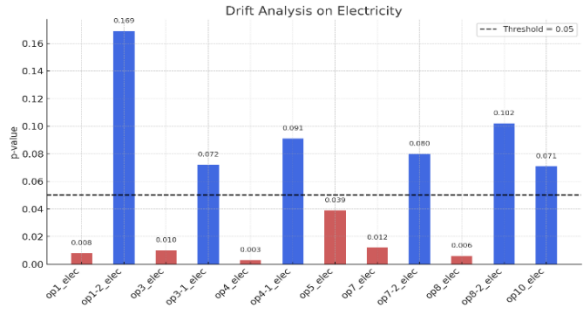


Fig. 3. Drift analysis (electricity).

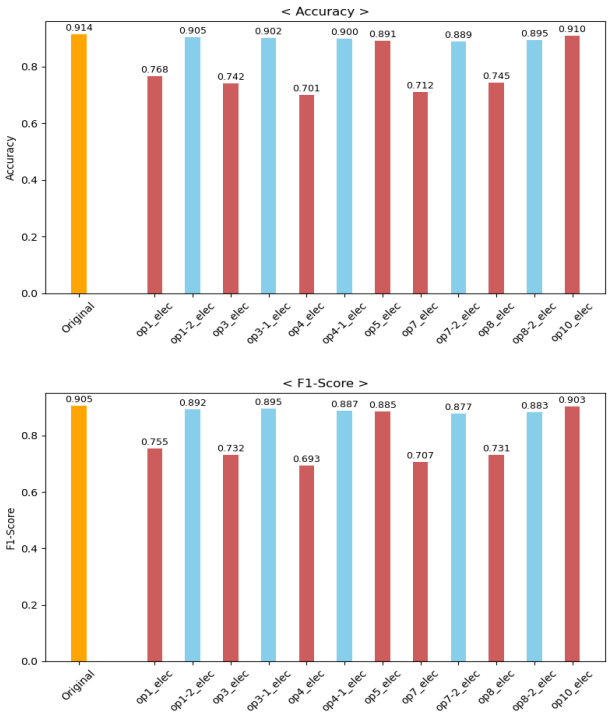


Fig. 4. Model performance (electricity).

C. Case II: CIFAR-10

1) Operational datasets generation: Ten CIFAR-10 operational datasets (op1_CIFAR-10 to op10_CIFAR-10) were generated to simulate brightness distortion, noise injection, blurring, and label mismatches.

2) Domain rule evaluation: The domain rules of the CIFAR-10 dataset are defined in Table IX, based on human visual quality perception and ML sensitivity thresholds.

TABLE IX. DOMAIN RULES AND THRESHOLDS (CIFAR-10)

NO	Rule	Threshold	Description
DR1	Brightness	80 < brightness < 180	Image brightness level within the perceptible range
FDR2	Blur	Sharpness score < 100	Image sharpness sufficient for recognition
DR3	Noise	Noise level < 30	Maximum acceptable noise level
DR4	Label consistency	No corrupted label values	All labels must be valid class indices (0–9)

TABLE X. DOMAIN RULE EVALUATION (CIFAR-10)

Dataset	DR Violations	Judgment	New Dataset
op1_ CIFAR-10	DR1(Brightness): 18%	Cleansing	op1-1
op2_ CIFAR-10	DR3(Noise): 25%	**Discard**	-
op3_ CIFAR-10	DR2(Blur): 12%	Cleansing	op3-1
op4_ CIFAR-10	DR1: 5%, DR2: 17%	Cleansing	op4-1
op5_ CIFAR-10	DR4(Label mismatch): 30%	**Discard**	-
op6_ CIFAR-10	DR1: 2%, DR3: 10%	Keep	
op7_ CIFAR-10	DR2: 9%	Keep	
op8_ CIFAR-10	DR3: 4%, DR4: 21%	Cleansing	op8-1
op9_CIFAR-10	DR1: 0%, DR2: 0%	Keep	
op10_CIFAR-10	DR1: 6%, DR2: 6%, DR3: 6%	Cleansing	op10-1

3) Data quality assessment: For CIFAR-10, perceptual quality metrics appropriate to 32×32 low-resolution images are applied (Table XI). PSNR, SSIM, NIQE, and BRISQUE thresholds are calibrated to the native resolution of CIFAR-10 images, ensuring that the quality criteria reflect realistic expectations for the target modality rather than high-resolution benchmarks (Table XII).

TABLE XI. QUALITY-SPECIFIC DATA REQUIREMENT (CIFAR-10)

NO	Quality Metric	Threshold	Description
DQ1	PSNR	≥ 28 dB	Peak Signal-to-Noise Ratio (applicable to 32×32 low-res images with adjusted baseline)
DQ2	SSIM	≥ 0.9	Structural Similarity Index
DQ3	NIQE	≤ 5.0	Naturalness Image Quality Evaluator
DQ4	BRISQE	≤ 40	Blind/Referenceless Image Spatial Quality Evaluator

TABLE XII. DATA QUALITY ASSESSMENT (CIFAR-10)

Dataset	DQ1	DQ2	DQ3	DQ4	Judgment	New Dataset
op1-1	26.5	0.91	5.5	45	Cleansing	op1-2
op3-1	29.1	0.93	4.9	39	Keep	-
op4-1	27.0	0.9	5.3	43	Cleansing	op4-2
op6	30.2	0.95	3.5	30	Keep	-
op7	31.5	0.96	3.2	28	Keep	-
op8-1	24.7	0.88	6.2	52	Cleansing	op8-2
op9	32.1	0.97	2.8	25	Keep	-
op10-1	25.9	0.89	5.9	47	Cleansing	op10-2

4) Drift analysis: FID (Fréchet Inception Distance) was applied to datasets passing the quality assessment step (Fig. 5). A threshold of FID > 40 denotes significant drift. Refined datasets (e.g., op1-2_CIFAR-10) exhibited substantially reduced FID compared to their unrefined counterparts (e.g., op1_CIFAR-10).

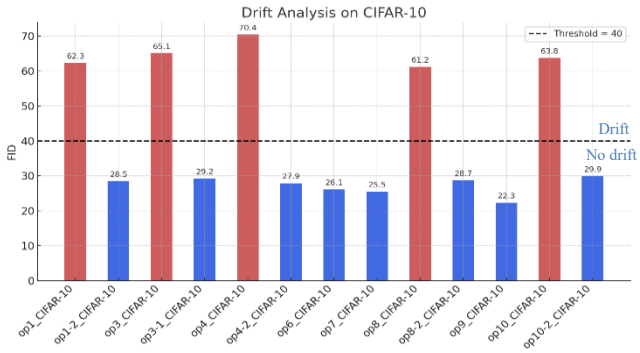


Fig. 5. Drift analysis (CIFAR-10).

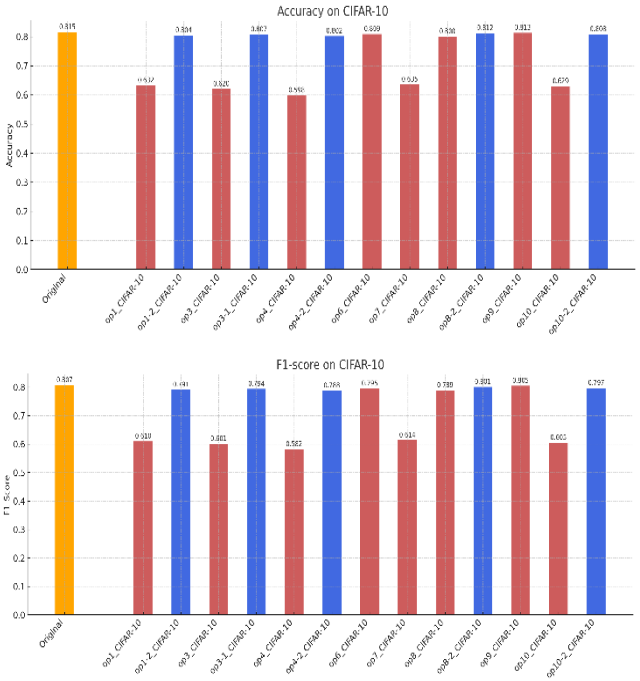


Fig. 6. Model performance (CIFAR-10).

5) Model performance comparison: Fig. 6 shows the model performance on the CIFAR-10 operational datasets. Average accuracy improved from 0.723 to 0.770, and F1-score improved from 0.726 to 0.774, representing improvements of 6.50% and 6.61%, respectively.

D. Case III: VisDA-2017

1) Operational dataset generation: Ten synthetic variants of the VisDA-2017 dataset (op1_Vis to op10_Vis) were generated to reflect image resolution shifts, channel misalignment, brightness distortion, and label noise.

2) Domain rule evaluation: Domain rules: DR1 (Image resolution), DR2 (RGB channel alignment), DR3 (Semantic label validity), DR4 (Brightness). Evaluation results are presented in Table XIII.

3) Data quality assessment: The same image quality metrics as CIFAR-10 (Table X) were applied, adjusted to VisDA-2017's higher-resolution inputs. Results are presented in Table XIV.

TABLE XIII. DOMAIN RULES EVALUATION (VISDA-2017)

Dataset	DR Violations	Judgment	New Dataset
op1_Vis	DR1(Resolution): 16%, DR2(Channels): 0%	Cleansing	op1-1_Vis
op2_Vis	DR3(Label mismatch): 28%	**Discard**	-
op3_Vis	DR4(Brightness): 11%	Cleansing	op3-1_Vis
op4_Vis	DR1: 5%, DR4: 21%	Cleansing	op4-1_Vis
op5_Vis	DR2: 12%	**Discard**	-
op6_Vis	DR1: 0%, DR4: 0%	Keep	op6_Vis
op7_Vis	DR3: 18%	Cleansing	op7-1_Vis
op8_Vis	DR4: 26%	Cleansing	op8-1_Vis
op9_Vis	DR2: 5%	Keep	op9_Vis
op10_Vis	DR1: 3%, DR2: 6%, DR3: 4%	Cleansing	op10-1_Vis

TABLE XIV. DATA QUALITY ASSESSMENT (VISDA-2017)

Dataset	DQ1	DQ2	DQ3	DQ4	Judgment	New Dataset
op1-1_Vis	27.4	0.89	5.2	44	Cleansing	op1-2_Vis
op3-1_Vis	25.1	0.87	5.9	49	Cleansing	op3-2_Vis
op4-1_Vis	26.5	0.88	5.4	47	Cleansing	op4-2_Vis
op6_Vis	29.8	0.92	3.7	31	Keep	
op7-1_Vis	26.2	0.89	5.6	46	Cleansing	op7-2_Vis
op8-1_Vis	24.0	0.85	6.4	53	Cleansing	op8-2_Vis
op9_Vis	30.1	0.93	3.2	28	Keep	
op10-1_Vis	25.8	0.86	5.7	50	Cleansing	op10-2_Vis

4) Drift analysis: FID was employed to detect distributional drift, with a threshold of FID > 40 denoting significant drift. Refined datasets demonstrated consistently lower FID scores compared to their unrefined counterparts (Fig. 7).

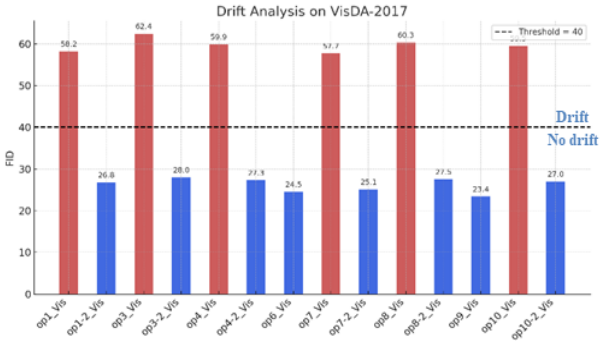


Fig. 7. Drift analysis (VisDA-2017).

5) Model performance comparison: By applying domain rule validation and quality assessment, accuracy improved from 0.568 to 0.639 and F1-score from 0.574 to 0.642, corresponding to improvements of 12.5% and 11.85%, respectively—the largest gains across all three datasets, reflecting the challenging synthetic-to-real distribution gap that DcDM's semantic validation effectively addresses (Fig. 8).

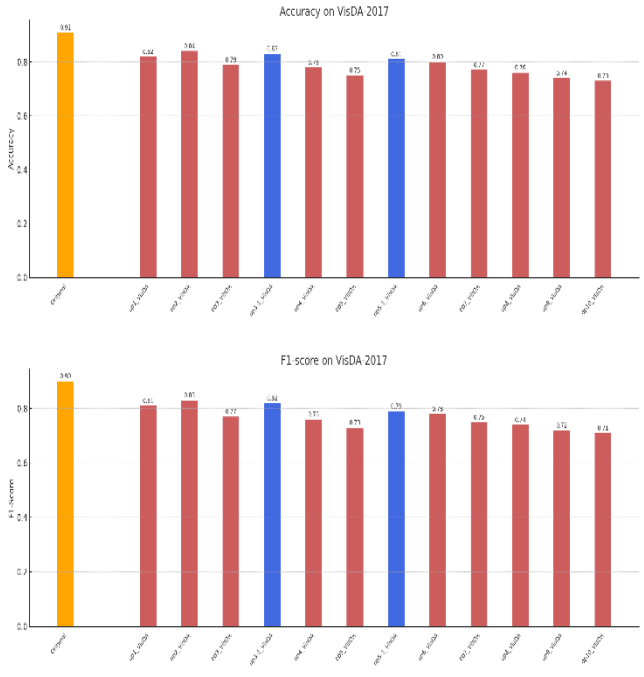


Fig. 8. Model performance (VisDA-2017).

E. Result Analysis

1) Statistical significance analysis: Paired t-tests were applied to the accuracy and F1-scores between refined and unrefined datasets. Across all three datasets, improvements were statistically significant at the 1% level ($p < 0.01$). Table XV summarizes the statistical significance of each dataset.

TABLE XV. STATISTICAL SIGNIFICANCE OF PERFORMANCE IMPROVEMENTS

Dataset	Metric	Baseline (Mean ± CI)	DcDM Refined (Mean ± CI)	p-value
Electricity	Accuracy	0.821 ± 0.012	0.869 ± 0.010	< 0.01
	F1-score	0.824 ± 0.013	0.872 ± 0.009	< 0.01
CIFAR-10	Accuracy	0.723 ± 0.015	0.770 ± 0.014	< 0.01
	F1-score	0.726 ± 0.016	0.774 ± 0.013	< 0.01
VisDA-20	Accuracy	0.568 ± 0.020	0.639 ± 0.018	< 0.01
	F1-score	0.574 ± 0.019	0.642 ± 0.017	< 0.01

Effect sizes (Cohen's $d > 0.6$) confirm that DcDM consistently yields moderate-to-strong gains over baseline methods, reinforcing the reliability of the reported results.

2) Ablation study: To isolate the contribution of each DcDM component, we conducted a controlled ablation study on the Electricity and CIFAR-10 datasets. We evaluated five configurations: no refinement (baseline), domain rule validation only, quality assessment only, domain rule + quality assessment (without drift detection), and the full DcDM pipeline. Results are presented in Table XVI.

The ablation results confirm that each component contributes incrementally to overall performance. Domain rule validation alone yields a moderate improvement of approximately 2.2–2.5% in accuracy by removing semantically invalid samples. Quality assessment alone provides a slightly larger gain (approximately 3.0–3.4%) by filtering low-quality data. Combining both without drift detection achieves near-full performance recovery, with an improvement of approximately 3.7–4.0%. The full DcDM pipeline, which additionally incorporates drift-triggered retraining, achieves the largest gains, confirming that all three components are necessary for optimal performance.

TABLE XVI. ABLATION STUDY: CONTRIBUTION OF EACH PIPELINE COMPONENT

Configuration	Electricity		CIFAR-10		VisDA	
	Acc.	F1	Acc.	F1	Acc.	F1
No refinement (Baseline)	0.821	0.824	0.723	0.726	0.568	0.574
Domain Rule Only	0.843	0.846	0.748	0.751	0.592	0.595
Quality Assessment Only	0.851	0.854	0.757	0.760	0.608	0.611
Domain Rule + Quality (DcDM without drift detection)	0.858	0.861	0.763	0.767	0.624	0.627
Full DcDM (All three steps)	0.869	0.872	0.770	0.774	0.639	0.642

3) Comparison with existing drift detection baselines: To contextualize DcDM's performance gains, we compared it against three established drift detection approaches: ADWIN [19], DDM [16], and Evidently AI [27], applied to the Electricity and CIFAR-10 datasets. All baselines operate reactively (detecting drift post-deployment) and do not incorporate semantic rule validation or data quality gating prior to training. Results are presented in Table XVII.

TABLE XVII. COMPARISON WITH REACTIVE DRIFT DETECTORS

Method \ Dataset	Electricity		CIFAR-10	
	Acc.	F1	Acc.	F1
No refinement	0.821	0.824	0.724	0.726
ADWIN	0.834	0.836	0.737	0.739
DDM	0.839	0.841	0.741	0.743
Evidently AI	0.845	0.847	0.748	0.751
DcDM (Ours)	0.869	0.872	0.770	0.774

The comparison should not be interpreted as a direct algorithmic competition because ADWIN, DDM, and Evidently AI were originally designed as reactive post-deployment monitoring solutions. The purpose of this comparison is to illustrate the practical difference between proactive data-centric prevention and reactive drift detection paradigms. ADWIN and DDM, which focus on detecting concept drift in streaming data without upstream data quality intervention, achieve modest improvements over unrefined data (approximately 1.3–1.8% accuracy gain). Evidently, AI, which monitors distributions at the feature level, provides slightly larger gains (approximately 2.4–2.7%), but still falls short of DcDM's 5.8–6.5% improvements. These results confirm that DcDM's proactive, pre-training integration of semantic validation and quality

assessment provides complementary and superior benefits over reactive monitoring approaches.

4) Sensitivity analysis on thresholds: The sensitivity analysis on thresholds (α for domain rules, β for quality metrics, γ for drift detection) is presented in Table XVIII. Performance remained stable with accuracy and F1-score fluctuating within ±2% under ±10% threshold variation, demonstrating robustness to threshold misconfiguration. Under extreme threshold variations of ±30%, performance degradation remained bounded below 6%.

TABLE XVIII. SENSITIVITY ANALYSIS ON THRESHOLDS

Threshold Setting	Accuracy (Mean)	F1-score (Mean)	Observation
α=5%, β=5%, γ=0.05	0.861	0.865	Conservative: fewer data passes; safe but slightly lower coverage
α=10%, β=10%, γ=0.05	0.869	0.872	Optimal balance: best accuracy/F1 combination across all datasets
α=20%, β=20%, γ=0.1	0.854	0.858	Slight degradation: more borderline data allowed through
α=30%, β=30%, γ=0.1	0.835	0.840	Clear degradation: lenient thresholds reduce filtering effectiveness

5) Overhead and computing cost: All experiments were conducted on a local workstation equipped with an Intel Core i7 processor, 32 GB of RAM, and an NVIDIA RTX 3080 GPU. Average processing time per step is summarized in Table XIX.

TABLE XIX. OVERHEAD AND COMPUTING COST

Dataset	Domain Rule Validation	Quality Assessment	Drift Detection
Electricity (100K records)	2.1 sec	3.4 sec	1.9 sec
CIFAR-10 (50K images)	3.2 sec	4.8 sec	2.7 sec
VisDA-2017 (50K images)	3.5 sec	5.0 sec	3.0 sec

Batch-level processing time for the Electricity dataset: domain rule validation $\approx$ 12 ms per 1,000 records, quality assessment ≈ 18 ms, drift detection (KS test) ≈ 9 ms. For image datasets, FID-based drift detection averaged 0.04 sec per batch of 500 images. These results indicate minimal computational overhead suitable for batch processing. For large-scale streaming environments, a two-level validation strategy is envisioned: 1) lightweight domain rules and quality checks in real time, and 2) comprehensive drift analysis asynchronously in batch mode.

We also measured batch-level processing time to evaluate the practicality of the framework. For the Electricity dataset, domain rule validation required approximately 12 ms per batch of 1,000 records, quality assessment required 18 ms, and drift detection (KS test) required 9 ms. Similar magnitudes were observed for image datasets, where FID-based drift detection averaged 0.04 sec per batch of 500 images, which corresponds to the reported totals for 50K images. These results indicate that the framework introduces minimal computational overhead and is therefore feasible for batch processing. For large-scale streaming data, we envision a two-level validation strategy: 1) lightweight domain rules and quality checks applied in real time at the high-velocity streaming environments, and 2) comprehensive validation performed asynchronously in batch mode.

VI. DISCUSSION

The DcDM framework demonstrates notable improvements in robustness and reliability by proactively addressing data drift. The main contributions include the integration of domain validation, statistical drift analysis, and automation in machine learning pipelines. Nevertheless, this study has several limitations.

1) Use of synthetic drift generation: All operational datasets were generated via synthetic perturbations. While this enables controlled experimentation, synthetic drift may not fully capture the complexity of real-world drift phenomena, such as non-linear, gradual, or concept-intertwined drift in industrial sensor networks or healthcare data streams. Future validation on real deployment logs or domain-specific streaming datasets (e.g., medical imaging or financial transactions) is necessary to confirm robustness under realistic drift conditions.

2) Reliance on domain experts for rule definition: The current framework relies on domain experts to define semantic rules, which can be costly and domain-specific. Automated rule induction through data profiling, pattern mining, or meta-learning could complement expert-driven rules. For example, profiling tools can automatically infer valid ranges, correlations, and label consistency from historical data, reducing human dependency in rapidly evolving domains.

3) Fixed thresholds: Although the sensitivity analysis confirmed robustness under moderate variation, fixed thresholds may not adapt optimally across all deployment scenarios. A promising direction is feedback-driven or reinforcement learning-based threshold adaptation, allowing DcDM to dynamically balance false alarms and missed drift detections as data distributions evolve.

4) No validation against real deployment environment: Although real-world deployment datasets with verified drift annotations were not available for this study, we attempted to approximate realistic operational conditions by applying perturbations derived from commonly reported degradation patterns, including sensor noise, missing values, illumination changes, image blur, and label inconsistencies. These perturbations represent typical causes of data-quality degradation reported in industrial monitoring, computer vision, and data-streaming applications. Future work will evaluate DcDM using naturally evolving datasets collected from production environments to further validate its practical effectiveness.

VII. CONCLUSION AND FUTURE WORK

This study proposed a proactive, data-centric methodology, DcDM (Data-centric Drift Management), for managing data drift across the ML lifecycle. Unlike traditional model-centric approaches that respond to performance degradation, DcDM

emphasizes early intervention by evaluating datasets before model training. The proposed framework systematically integrates three components: 1) domain rule validation to assess semantic correctness, 2) quality assessment based on goal-driven metrics, and 3) statistical drift detection to quantify distributional shifts.

By applying this methodology across diverse datasets—Electricity Load Diagrams, CIFAR-10, and VisDA-2017—we demonstrated performance improvements of 6%–12.5% in both accuracy and F1-score. An ablation study confirmed that each pipeline component contributes incrementally to overall gains, and a baseline comparison against ADWIN, DDM, and Evidently AI confirmed that DcDM's proactive, pre-training integration provides superior benefits over reactive monitoring approaches. Statistical significance tests ($p < 0.01$, Cohen's $d > 0.6$) and sensitivity analyses confirm the robustness of these results.

Future work will focus on: 1) validating DcDM on real-world deployment logs and natural drift scenarios; 2) automating domain rule induction via data profiling and meta-learning to reduce expert dependency; 3) developing feedback-driven, adaptive threshold mechanisms; 4) extending DcDM to explicitly handle concept drift; and 5) integrating DcDM into production-grade MLOps pipelines with privacy-preserving rule management.

ACKNOWLEDGMENT

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2026-RS-2022-00156334).

REFERENCES

[1] C. Zhang, Y. Bengio, and M. Hardt, “Understanding deep learning requires rethinking generalization,” arXiv preprint arXiv.1611.03530, https://doi.org/10.48550/arXiv.1611.03530.

[2] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A survey on concept drift adaptation,” ACM Comput. Surv., vol. 46, no. 4, pp. 44:1–44:37, Mar. 2014, doi: 10.1145/252381.

[3] A. Tsymbal, “The problem of concept drift: Definitions and related work,” Computer Science Department, Trinity College Dublin, Tech. Rep. TCD-CS-2004-15, 2004.

[4] R. Baier, F. Heine, and F. Puppe, “Handling concept drift in process mining,” Expert Systems with Applications, vol. 128, pp. 60–77, 2019.

[5] S. Wang, M. Minku, and X. Yao, “A review of online class imbalance learning: Trends and challenges,” Progress in Artificial Intelligence, vol. 5, no. 3, pp. 123–135, 2016.

[6] A. Webb, D. Page, and T. A. Miller, “A real-time approach to concept drift detection and model refresh in streaming data,” in Proc. ACM Int. Conf. Information and Knowledge Management (CIKM), 2020.

[7] A. Breck, S. Cai, E. Nielsen, J. Salib, and D. Sculley, “The ML test score: A rubric for ML production readiness and technical debt reduction,” in Proc. SysML Conf., 2019.

[8] A. Bontempelli et al., “Data-centric AI: A comparative review,” Data-Centric AI Workshop at NeurIPS, 2021.

[9] A. Zoller, and M. Huber, “Benchmarking the robustness of machine learning models to data quality issues,” arXiv preprint, arXiv:2202.04643, 2022.

[10] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, “Learning under concept drift: A review,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 12, pp. 2346–2363, Dec. 2019.

[11] Moez Ali, “Understanding data drift and model drift: Drift detection in Python,” DataCamp, 2023. [Online]. Available: https://www.datacamp.com/tutorial/understanding-data-drift-model-drift. [Accessed: Jun. 2025].

[12] B. Yurdakul and J. Naranjo, "Statistical properties of populations stability index", Journal of Risk Model Validation, vol. 14, no. 4, pp.89-100, 2020.

[13] S. Kullback, and R. A. Leibler, “On information and sufficiency”, The Annals of Mathematical Statistics, vol. 22, no. 1, pp.79-86, 1951.

[14] F. Nielsen, “On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid”, Entropy, vol. 22, no. 2, Art. No. 221, 2020.

[15] S. Ashok, S. Ezhumalai, T. Patwa, "Remediating data drift and re-establishing ML models", Procedia Computer Science, vol. 218, pp.799-809, 2023.

[16] I. Žliobaitė, “Learning under concept drift: Early drift detection method,” *in* Lect. Notes Comput. Sci. (LNAI 3171), pp. 286–295, 2004, https://doi.org/10.1007/978-3-540-28645-5_29.

[17] Manuel Baena-García, José Campo-Ávila, Raúl Fidalgo-Merino, Albert Bifet, Ricard Gavald, and Rafael Morales-Bueno. “Early drift detection method,” in Proceedings of the International Workshop on Knowledge Discovery from Data Streams, pp.77–86. 2006.

[18] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, “Exponentially weighted moving average charts for detecting concept drift,” Pattern Recognit. Lett., vol. 33, no. 2, pp. 191–198, Dec. 2012, https://doi.org/10.48550/arXiv.1212.6018

[19] A. Bifet and R. Gavaldà, “Learning from time-changing data with adaptive windowing,” in Proc. SIAM Int. Conf. Data Mining, 2007, pp. 443–448, https://doi.org/10.1137/1.9781611972771.42

[20] I. Frías-Blanco, J. d. Campo-Ávila, G. Ramos-Jiménez, R. Morales-Bueno, A. Ortiz-Díaz, and Y. Caballero-Mota, “Online and non-parametric drift detection methods based on Hoeffding’s bounds,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 810–823, Mar. 2015, https://doi.org/10.1109/TKDE.2014.2345382

[21] A. Gretton, et al., “A kernel two-sample test”. Journal of Machine Learning Research, vol. 13, pp.723-773, 2012.

[22] Min Jin Chong and David Forsyth, “Effectively unbiased FID and inception score and where to find them”, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit(CVPR) 2020, https://doi.org/10.48550/arXiv.1911.07023

[23] B. Celik and J. Vanschoren, “Adaptation strategies for automated machine learning on evolving data”, IEEE Transactions on Pattern Analysis and Machine Intelligence“, vol.43, no.9, pp. 3067-3078, 2021, https://doi.org/10.1109/TPAMI.2021.3062900

[24] E. Kalaitzidis et al., “AML4S: An AutoML pipeline for data data streams”, MDPI Machine learning & Knowledge extraction, 7(3), pp1-33, July. 2025, https://doi.org/10.3390/make7030087

[25] M. Schaller et al., “AutoML for multi-class anomaly compensation of sensor drift”, arXiv preprint arXiv:2502.19180, 2025, https://doi.org/10.48550/arXiv.2502.19180

[26] A. Ng, "A chat with Andrew on MLOps: From model-centric to data-centric AI", DeepLearningAI, 2021, [Online]. Available: https://duly.kr/4bfuMF5

[27] E. Breck et al., "Evidently AI: Open-source ML monitoring and testing," GitHub repository, 2022. [Online]. Available: https://github.com/evidentlyai/evidently

[28] J. Van Looveren et al., "Alibi Detect: Algorithms for outlier, adversarial and drift detection," J. Mach. Learn. Res., 2022.

[29] UCI Electricity Load Diagrams Dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams. [Accessed: Jun. 2025].

[30] CIFAR-10 Dataset. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html. [Accessed: Jun. 2025]

[31] VisDA-2017 Dataset. [Online]. Available: https://github.com/VisionLearningGroup/visda-2017 [Accessed: Jun. 2025]

Characterizing Operational Drift in Cement Manufacturing Process Data via Self-Supervised Representation Learning

A Multi-Year Operational Analysis Using PELT Change-Point Detection and SCARF Embeddings

Changgyun Kim
Department of Advanced AI Engineering, Samcheok, South Korea

***Abstract*—Cement finish milling generates large volumes of process variable data at hourly resolution, while quality measurements (Blaine fineness, 44 μm residue) are recorded only every two to four hours, producing a sparse-label regime with substantial unlabeled data accumulated over multi-year operation. In this study, we analyze 188,858 hourly records collected from three industrial cement mill units spanning 2017 to 2025 (approximately nine years) and characterize the operational drift structure embedded in the data. Using Pruned Exact Linear Time (PELT) change-point detection on three critical process variables, we identify between 12 and 21 detected drift events per mill, with Cohen's d effect sizes exceeding 1.0 for the majority of detected change points and reaching 4.3 in extreme cases. We then apply SCARF, a contrastive self-supervised learning method for tabular data, to learn 128-dimensional representations on the combined labeled (51,225 records) and unlabeled (111,506 records) data. Multi-seed training yields stable validation InfoNCE loss of 6.93 ± 0.02. Three clustering algorithms applied to the learned embedding space (K-means, Gaussian Mixture Model, HDBSCAN) consistently select 15 to 21 operational regimes, with silhouette scores between 0.36 and 0.41. The Adjusted Rand Index between embedding-space K-means and process variable space K-means is 0.35, indicating that the learned representation preserves coarse regime structure while resolving finer sub-regime variability. Cluster analysis further reveals strong mill specificity, with 14 of 15 embedding clusters dominated by a single mill, and temporal cluster evolution that aligns with PELT-detected change-point boundaries. These findings establish that long-term cement process data contains a richer operational regime structure than implied by raw process variable clustering, and that self-supervised pretraining can recover this structure, and that the resulting representation yields statistically significant gains in downstream quality prediction.**

***Keywords*—Cement manufacturing; self-supervised learning; change-point detection; operational drift; tabular data; representation learning; SCARF; PELT**

I. Introduction

Cement manufacturing is a globally significant industrial process responsible for an estimated 5 per cent of anthropogenic carbon dioxide emissions [1]. The finish milling stage, in which clinker is ground together with gypsum and supplementary cementitious materials, determines the principal product quality indices: Blaine fineness (cm²/g) and the residue retained on a 44 μm sieve (44 μmR, %). Stable control of these indices is essential because deviations propagate directly into customer-facing product specifications and into the energy efficiency of subsequent hydration and strength development.

Industrial cement facilities typically collect hourly process variable (PV) measurements from distributed control system (DCS) sensors, producing dense multivariate time series at the input side. However, quality measurements depend on laboratory sampling and are recorded only every two to four hours. This temporal asymmetry yields a structurally sparse-label regime in which only a small fraction of process records can be directly paired with quality labels for supervised learning. Recent benchmarks in the tabular learning literature [13] have reported that gradient-boosted tree ensembles remain competitive with or superior to state-of-the-art deep tabular architectures at sample sizes typical of industrial sparse-label settings, a pattern also observed in preliminary analysis of this facility's data.

Initial inspection of the full dataset revealed that pre-2023 Blaine measurements exhibit substantial distributional deviation relative to post-2023 records. Restricting analysis to the 2023 to 2025 period to avoid these artifacts reduced the available process records by approximately 87 percent (from 188,858 to 24,878) and limited the data span to roughly three years. The present analysis revisits this design choice by treating the pre-2023 period as a candidate source of unlabeled and partially labeled data, applying a wider valid-range filter that retains those pre-2023 records whose Blaine and 44 μmR fall within plausible operational bounds while excluding records exhibiting clear out-of-range artifacts. This study then investigates whether the retained pre-2023 records, combined with all available 2023 to 2025 records, contain a coherent operational regime structure that can be recovered through self-supervised representation learning.

The contributions of this study are threefold. First, we apply Pruned Exact Linear Time (PELT) change-point detection [2] with automatic penalty selection to three critical process variables across all three mills and characterize the temporal drift structure embedded in approximately nine years of operational data. Second, we apply SCARF [3], a contrastive self-supervised learning method specifically designed for tabular data, to learn 128-dimensional representations on the

combined labeled and unlabeled data. Third, we conduct a systematic clustering analysis on the learned embedding space using K-means, Gaussian Mixture Models, and HDBSCAN, and validate the resulting regimes against the change-point structure obtained independently through PELT.

The remainder of this study is organized as follows: Section II reviews related work on cement quality modeling, self-supervised learning for tabular data, and change-point detection. Section III describes the dataset and preprocessing pipeline. Section IV characterizes operational drift using PELT and K-means clustering on raw process variables. Section V describes the SCARF self-supervised pretraining method. Section VI presents embedding-space clustering and interprets the discovered regime structure with respect to mill identity and temporal evolution. Section VII discusses implications for downstream quality prediction and limitations of the present study. Section VIII concludes.

II. Related Work

A. Quality Prediction in Cement Manufacturing

Data-driven modeling in the cement and process industries has been surveyed extensively. Ge et al. [4] reviewed machine learning techniques applied across process-industry tasks, organizing approaches by supervision regime and identifying soft-sensor modeling as one of the principal application areas. Yuan et al. [5] introduced a deep stacked autoencoder with variable-wise weighting for soft-sensor modeling in nonlinear dynamic industrial processes. Pani and Mohanta [6] applied support vector regression, fuzzy inference, and adaptive neuro-fuzzy inference for online monitoring of cement fineness. Yuan et al. [7] developed a weighted linear dynamic system for feature representation in soft-sensor applications across nonlinear dynamic industrial processes. Zermane and Drardja [8] developed a random forest based monitoring system for cement production. Although the cited soft-sensor and process-industry studies do not directly compare tree-based and deep alternatives, recent benchmarks in the broader tabular learning literature [13] have reported that gradient-boosted tree ensembles remain competitive with or superior to the state-of-the-art deep tabular architectures at the sparse-label sample sizes typical of cement finish-milling operations.

B. Self-Supervised Learning for Tabular Data

Self-supervised learning (SSL) for tabular data has emerged as a methodology for leveraging the abundant unlabeled records present in industrial process data. VIME [9] learns representations through value reconstruction and masking. SCARF [3] adopts a contrastive framework in which positive views are generated through feature corruption (random replacement from the marginal distribution) and the encoder is trained with InfoNCE loss [10]. SubTab [11] partitions feature subsets and learns through cross-subset reconstruction. TabTransformer [12] applies attention-based encoders to tabular inputs. Recent benchmarks [13] show that SSL pretraining can narrow the performance gap between deep tabular models and gradient-boosted trees when downstream labeled data is limited, although these results have not been systematically validated on industrial datasets exhibiting operational drift, which is the setting of the present study.

C. Change-Point Detection and Operational Drift Identification

Change-point detection [14] aims to identify time points, where the statistical properties of a signal change. PELT [2] achieves linear time complexity through optimal pruning, with the segmentation penalty parameter controlling the number of detected change points. For multivariate signals or single signals with mean-shift transitions, the L2 cost model is commonly applied. Drift identification in industrial process monitoring has traditionally relied on statistical process control charts [15]; more recent approaches use machine learning-based concept drift detection [16]. In this study, we apply PELT with automatic penalty selection to recover the operational drift structure in cement process variables.

III. Dataset and Preprocessing

A. Data Source

The dataset consists of DCS records from three cement finish milling units, designated CM2, CM3, and CM4, at a commercial cement facility in the Republic of Korea, covering the period from January 2017 to October 2025. The raw dataset comprises 188,858 hourly records, each associated with 39 process variables (mill drive load, separator load, bag filter differential pressure, ball mill power, feed rates of clinker, gypsum, slag and fly ash, classifier and separator rotational speeds, and auxiliary roller and pressure measurements) and two quality target variables (Blaine fineness and 44 μmR). All variables were already aggregated to hourly resolution by the facility's historian system.

B. Labeled and Unlabeled Partitioning

Quality measurements at the finish milling stage are recorded approximately every two to four hours, while process variables are sampled hourly. To construct a valid labeled set, we apply a backward-aligned causal pairing rule: a process record at time t is paired with the most recent prior or simultaneous laboratory quality measurement that satisfies all of the following conditions: (1) the time gap is at most three hours, (2) Blaine fineness falls within [2,500, 5,000] cm^2/g, and (3) 44 μmR falls within [0, 25]%. Records that fail either quality range condition are excluded from the labeled set rather than being relegated to the unlabeled pool, because their associated quality values would otherwise contaminate downstream supervised training. The unlabeled set comprises records for which no laboratory quality measurement exists within the causal window.

This partitioning produces a labeled set D_1 of 51,225 records and an unlabeled set D_unlab of 111,506 records (see Table I). Relative to a period-restricted preliminary analysis of the same facility data, D_1 expands the labeled set through two mechanisms operating jointly: (1) the wider valid-range filter applied uniformly across the full 2017 to 2025 period admits pre-2023 records that pass both Blaine and 44 μmR range conditions, and (2) the wider range admits additional 2023 to 2025 records that were excluded under the stricter range. Approximately two-thirds of the increase in D_1 originates from the 2023 to 2025 period, while approximately one-third originates from pre-2023 records that satisfy both range conditions. The remaining pre-2023 records that fail either range

condition are excluded from both D_1 and D_unlab; only records lacking quality labels enter D_unlab.

The labeled set D_1 is partitioned into training (35,854 records, 70 per cent), validation (7,680 records, 15 per cent), and test (7,691 records, 15 per cent) subsets using a regime-stratified random split. A K-means clustering is fitted on D_1 in the 40-dimensional space formed by the 37 standardized process variables described in Section III C, concatenated with three mill one-hot indicator dimensions, with the number of clusters automatically selected by silhouette score over $k \in \{3, \ldots, 8\}$, yielding $k = 8$. A 70/15/15 random split is then performed within each of the eight clusters to produce regime-balanced train, validation, and test subsets. The same eight clusters are reused as the raw-PV reference clustering in Section IV B, so that the comparison between raw-PV and embedding-space clusterings is based on a single, fixed PV-space partition rather than two independently fitted K-means models. This protocol is adopted to ensure regime-balanced evaluation across the full 2017 to 2025 data span and to allow future comparison with downstream supervised tasks using the same partition. The fitted K-means model is subsequently applied to D_unlab via nearest-centroid assignment to provide regime labels for all 162,731 records used in the ARI computation of Section VI. We acknowledge that the random subset selection from a multi-year time-series introduces residual temporal contiguity between training and test records and is, in this sense, weaker than a strictly chronological split for assessing temporal generalization; this trade-off is discussed further in Section VII D.

C. Process Variable Preprocessing

From the 39 process variables, two with greater than 50 per cent missing values across the full 2017 to 2025 period are removed (POLYCOM_IMPELLER_CRUSHER and ETC_GRINDING_AID), yielding 37 retained PVs. Records with more than 50 per cent of PVs missing are treated as plant shutdown periods and are excluded from both D_1 and D_unlab. Remaining missing values within retained records are imputed using per-mill forward fill followed by backward fill. Multivariate outliers are then filtered using IsolationForest with contamination 0.02 and 200 trees, fitted jointly on the combined $D_1 \cup$ D_unlab feature matrix. The remaining 37 PVs, together with the mill identifier (one-hot encoded over CM2, CM3, CM4), constitute the 40-dimensional input feature space for both representation learning and downstream tasks.

TABLE I. DATASET SUMMARY

Subset	Mills	Records	Time range	Purpose
D_1 train	CM2/3/4	35,854	2017–2025	Future supervised training
D_1 val	CM2/3/4	7,680	2017–2025	Hyperparameter selection
D_1 test	CM2/3/4	7,691	2017–2025	Held-out evaluation
D_unlab	CM2/3/4	111,506	2017–2025	SSL pretraining
Total	CM2/3/4	162,731	2017–2025	—

[a] Splits are regime-stratified random subsets of D_1; time range refers to the data span, not subset-specific intervals. Total raw records: 188,858. Difference comprises records dropped by shutdown filtering, IsolationForest outlier filtering, and quality value range filtering (records failing either Blaine or 44 μmR range conditions are excluded entirely).

IV. OPERATIONAL DRIFT CHARACTERIZATION

Fig. 1 provides an overview of the complete analysis pipeline. The raw multivariate process data, after the preprocessing described in Section III, form a 40-dimensional input consisting of 37 standardized process variables and three mill one-hot indicators. This input feeds two complementary analyses. First, PELT change-point detection is applied to selected critical signals to characterize the temporal operational drift directly in the process-variable space (Section IV-A). Second, the same input is used for SCARF contrastive self-supervised pretraining, in which a shared encoder f and projection head g are trained under an InfoNCE objective on corrupted and uncorrupted feature views to produce a 128-dimensional representation (Section V). The learned embedding is then clustered to recover operational regimes, which are validated against the independently obtained change-point structure (Section VI). The remainder of this section begins with the PELT analysis.

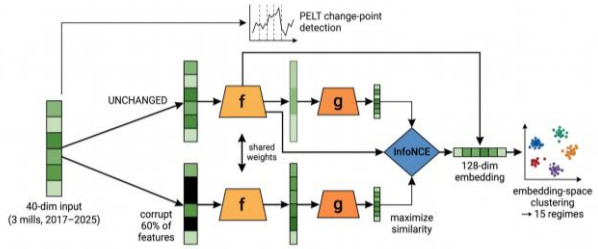


Fig. 1. Overview of the proposed pipeline (preprocessing, PELT drift detection, SCARF pretraining, and embedding-space clustering).

A. PELT Change-Point Detection

To characterize operational drift, we apply PELT change-point detection [2] to three operationally critical process variables: MILL_186_BF_PRESSURE (bag filter differential pressure, indicating grinding fineness), POLYCOM_WF_FEED_CLINKER (clinker feed rate, the dominant material input), and MILL_SEPOL_SEPOL (separator rotational speed). Each variable is aggregated to a daily resolution per mill, with a minimum of six valid hourly records required for valid daily aggregation. The L2 cost model is used with a minimum segment size of 60 days. The penalty parameter is selected automatically through grid search over {100, 300, 1,000, 3,000, 10,000, 30,000, 100,000, 300,000}, with a target range of 3 to 8 change points per signal per mill. When no penalty in the grid yields a change-point count within the target range, the penalty producing the count closest to the upper bound (8) is selected; this fallback applies to one signal-mill combination in the present analysis, CM4 POLYCOM_WF_FEED_CLINKER, which yields 10 change points at penalty 3,000 (the closest available value above the target upper bound). All other eight signal-mill combinations fall within the [3, 8] target range without fallback.

These three variables were selected to span the three principal and largely independent axes that govern finish-milling behavior rather than to exhaust the variable set: grinding fineness at the separation stage (MILL_186_BF_PRESSURE, the bag-filter differential pressure that reflects the attained product fineness), the dominant material input to the circuit

(POLYCOM_WF_FEED_CLINKER, the principal feed stream), and the classification control that sets the cut size (MILL_SEPOL_SEPOL, the separator rotational speed). This choice is further supported by an independent supervised quality-prediction analysis on data from the same facility, in which bag-filter differential pressure and separator-related variables emerge among the most influential predictors of both Blaine fineness and 44 μmR. Restricting the change-point analysis to these three complementary signals, therefore, targets the axes most likely to carry operationally meaningful drift while keeping the per-signal segmentation interpretable; a comprehensive multivariate change-point analysis over all 37 process variables is identified as future work in Section VII E.

Effect sizes are quantified using Cohen's d between adjacent segments, where $d = (\mu_2 - \mu_1) / s_{pooled}$ and the pooled standard deviation is computed across the two adjacent segments. By convention, $|d| > 0.8$ is interpreted as a large effect and $|d| > 1.2$ as very large. Table II summarizes the detected change points per mill and per process variable.

TABLE II. PELT CHANGE-POINT DETECTION RESULTS PER MILL

Mill	MILL_186_BF_PRESSURE	POLYCOM_WF_FEED_CLINKER	MILL_SEPOL_SEPOL
CM2	4 CPs (|d|max = 3.37)	3 CPs (|d|max = 1.50)	5 CPs (|d|max = 4.34)
CM3	7 CPs (|d|max = 2.93)	8 CPs (|d|max = 2.18)	6 CPs (|d|max = 2.85)
CM4	5 CPs (|d|max = 2.46)	10 CPs (|d|max = 2.05)	4 CPs (|d|max = 2.34)

[b.] CP = change point. |d|max denotes the maximum Cohen's d among all adjacent segment pairs for that variable and mill. The 10-CP count for CM4 POLYCOM_WF_FEED_CLINKER reflects the fallback rule described in the text.

The total number of detected change points per mill ranges from 12 (CM2) to 21 (CM3), with the majority of effect sizes satisfying $|d| > 1.0$ and several reaching $|d| > 3.0$. Importantly, the change points are distributed across all three process variables and all three mills rather than concentrated at a single temporal event, which would be the expected signature of a sensor recalibration or unit conversion failure. The pattern is instead consistent with progressive operational adjustments, equipment maintenance, and grinding aid composition changes that accumulate over multi-year operation. Fig. 2 illustrates the temporal pattern for MILL_186_BF_PRESSURE.

B. K-Means Regimes on Raw Process Variables

The eight K-means clusters introduced in Section III-B (fitted on D_1 in the 40-dimensional space of 37 standardized PVs plus three mill one-hot indicators) serve a second role in this section as the raw-PV reference clustering against which the embedding-space clusters in Section VI will be compared. Adopting a single, fixed clustering for both purposes avoids the ambiguity of comparing two independently fitted K-means models and ensures that the comparison in Section VI-B isolates the effect of the SCARF representation rather than any residual algorithmic variability. The silhouette score for this $k = 8$ partition is 0.16, computed on the same evaluation subset used for k selection. We note that $k = 8$ is the upper boundary of the search grid; while extension of the search range may yield higher silhouette values, the relatively low absolute silhouette score (a value of this magnitude indicates substantial cluster overlap, as silhouette scores below approximately 0.25 are commonly interpreted as evidence of weak cluster structure) reflects the overlapping nature of operating modes in the raw PV space. Fig. 2 overlays the K-means regime assignments as background shading, where each color corresponds to one regime.

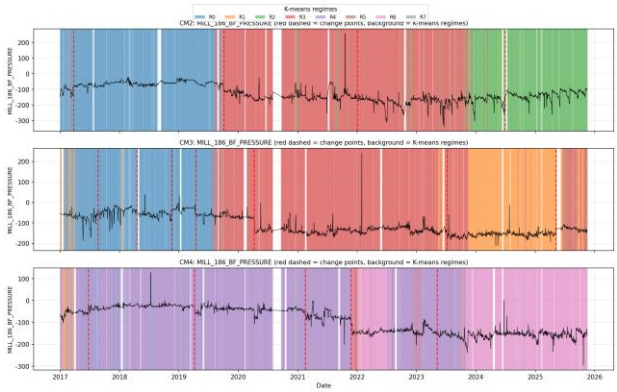


Fig. 2. MILL_186_BF_PRESSURE time series for CM2 (top), CM3 (middle), and CM4 (bottom). Red dashed lines: PELT change points. Background shading: K-means regimes on raw process variables.

Two observations from Fig. 2 are noteworthy. First, the K-means regime assignments are not temporally monotone: multiple regimes alternate across multiple time periods, consistent with operational cycling rather than a one-time global shift. Second, the K-means regime boundaries visually align with the PELT change-point locations, despite K-means having no temporal information. This convergence between the two independent methods (one purely temporal, one purely statistical in feature space) provides cross-method evidence that the discovered regimes correspond to genuine operational variability rather than algorithmic artifact.

V. SELF-SUPERVISED REPRESENTATION LEARNING

A. SCARF Architecture and Training

We adopt SCARF [3] for self-supervised representation learning on the combined dataset $D_1 \cup D_unlab$ (162,731 records). The architecture comprises an encoder f mapping the 40-dimensional input (37 standardized PVs concatenated with three mill one-hot dimensions) through a three-layer multilayer perceptron (MLP) with a hidden size of 256 to a 128-dimensional embedding, followed by a projector g consisting of a two-layer MLP (128 to 128) used only during pretraining. After pretraining, the projector is discarded, and only the encoder is used to produce embeddings.

SCARF training optimizes the InfoNCE contrastive loss [10], in which each sample's positive view is generated through feature corruption (random replacement of 60 per cent of features with values sampled from the empirical marginal distribution of each feature) and the embedding is trained to maximize similarity between corrupted positive pairs versus dissimilarity to other samples in the batch. The InfoNCE loss for sample i in a batch of size N is defined as in Eq. (1):

$$L_{NCE}(i) = -\log \left[\exp(\mathrm{sim}(z_i, z_i^+)/\tau) / \Sigma_{j \in B(i)} \exp(\mathrm{sim}(z_i, z_j)/\tau) \right] \quad (1)$$

where, $z = g(f(x))$ is the projected embedding, $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity, $\tau = 0.5$ is the temperature parameter, and $B(i)$ denotes the set of all $2N - 1$ distinct in-batch views excluding

the anchor z_i itself (that is, the positive view z_i^+ and all other corrupted views from the same mini-batch of size N = 1,024). The implementation follows the standard SimCLR-style 2N-batch convention. Training uses AdamW optimizer with learning rate 1×10^{-3} and weight decay 1×10^{-5}, batch size 1,024, and maximum 100 epochs with early stopping (patience 15) on validation InfoNCE loss measured on a 10 percent random held-out split of $D_1 \cup D_unlab$. The model is initialized from three random seeds (42, 123, 456), and the embedding with the lowest validation loss is selected for downstream analysis.

B. Pretraining Convergence

Multi-seed pretraining yields stable convergence behavior. The best validation InfoNCE losses across the three seeds are 6.92, 6.92, and 6.96, with a mean of 6.93 and a standard deviation of 0.02. For the 2N-batch convention used here with N = 1,024, the random-baseline expected loss under uniform negative sampling is $\log(2N - 1) = \log(2{,}047) \approx 7.62$ nats; the observed losses are below this baseline by approximately 0.7 nats. We acknowledge that the absolute InfoNCE value is sensitive to the choice of negative sampling convention, and the downstream evidence for meaningful representation learning is provided by the clustering quality and external validity analyses in Section VI rather than by the raw InfoNCE value alone. Fig. 3 shows training and validation loss curves across the three random seeds.

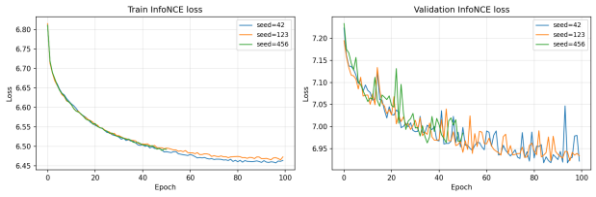


Fig. 3. SCARF training and validation InfoNCE loss curves across three random seeds.

VI. Embedding-Space Clustering

A. Clustering Algorithms and Selection

After pretraining, we apply L2 normalization to the 128-dimensional embeddings, projecting each sample onto the unit hypersphere. This is standard practice for contrastive embeddings, where cosine similarity (rather than Euclidean distance) was the training objective. Three clustering algorithms are then compared: K-means with silhouette-based k selection over {3, 5, 8, 10, 12, 15, 20}; Gaussian Mixture Model (GMM) with diagonal covariance and Bayesian Information Criterion (BIC) selection over {3, 5, 8, 10, 12, 15}; and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) with minimum cluster size selection over {200, 500, 1,000, 2,000} and minimum samples 50.

TABLE III. Embedding-Space Clustering Comparison

Algorithm	k	Silhouette	Davies–Bouldin	Calinski–Harabasz	ARI vs PV	Noise %
K-means	15	0.366	1.132	5,338	0.354	0.0%
GMM	15	0.363	1.072	5,245	0.378	0.0%
HDBSCAN	21	0.407	0.948	4,959	0.234	16.5%

[c.] Silhouette and Davies–Bouldin computed on a 30,000-sample subset for tractability. ARI is computed on full data.

Table III reports the clustering quality metrics. K-means selects k = 15 with silhouette 0.366, GMM selects k = 15 by BIC with silhouette 0.363, and HDBSCAN identifies 21 components at minimum cluster size 2,000 with silhouette 0.407 on non-noise samples (16.5 per cent of samples designated as noise). The Adjusted Rand Index (ARI) between each embedding-space clustering and the PV-space K-means clustering (Section IV) is 0.354 for K-means, 0.378 for GMM, and 0.234 for HDBSCAN. These ARI values indicate weak-to-moderate agreement, meaning that the embedding-space clusters preserve coarse PV-regime structure while introducing finer subdivisions. Fig. 4 visualizes the relative performance of the three algorithms across all four metrics.

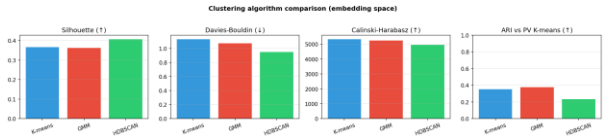


Fig. 4. Clustering algorithm comparison in the embedding space (silhouette, Davies–Bouldin, Calinski–Harabasz, and ARI versus PV K-means).

Three observations are worth emphasizing. First, the embedding-space silhouette score (0.37) substantially exceeds the raw PV-space K-means silhouette (0.16), indicating that the SCARF representation has produced a feature space where operating regimes are more cleanly separable; we acknowledge that the two silhouettes are computed in spaces of different dimensionality (128-dimensional L2-normalized embedding versus 40-dimensional standardized PV plus mill one-hot space) and that the contrastive objective directly optimizes for cluster-like compactness, so the quantitative ratio of the two silhouettes should be interpreted as a directional indicator rather than an equal-footing measurement. Second, K-means and GMM converge on the same number of components (k = 15) through independent selection criteria (silhouette and BIC, respectively). For K-means, k = 15 is an interior optimum of the wider search grid {3, 5, 8, 10, 12, 15, 20} (the silhouette score peaks at k = 15 and decreases at k = 20), which provides a stronger indication that k = 15 reflects an intrinsic count rather than a grid-boundary effect. For GMM, k = 15 coincides with the upper boundary of its search grid {3, 5, 8, 10, 12, 15}; BIC may favor an even larger number of components if the grid were extended, so the GMM result alone would be inconclusive. The agreement between the two algorithms — one finding an interior optimum and the other reaching its boundary at the same value — is consistent with k = 15 being an intrinsic count, though extending the GMM grid would strengthen the conclusion further. Third, HDBSCAN identifies 21 components with the highest silhouette, where the selected minimum cluster size of 2,000 was the largest value in its search range and yields 16.5 percent of samples designated as noise; for downstream pseudo-labeling applications this noise designation is operationally problematic. Based on these considerations and to ensure consistency with the visualizations in subsequent subsections, we adopt K-means with k = 15 as the primary representation for the remainder of this study.

To assess whether the contrastive representation is necessary, or whether a simpler and computationally cheaper pipeline would suffice, we additionally clustered the same 162,731-record input using linear dimensionality reduction. Table IV compares K-means (k = 15) applied to PCA-reduced

and standardized process-variable representations against the SCARF embedding. Two patterns emerge. First, the SCARF embedding produces markedly more separable clusters: its silhouette score (0.366) is roughly twice that of any PCA or standardized-PV baseline (0.16–0.24), indicating that linear projection preserves the diffuse, overlapping structure of the raw process-variable space rather than resolving it. Second, the PCA baselines exhibit higher ARI against the PV-space K-means partition (0.47–0.51 versus 0.354 for SCARF), which is expected: a linear projection largely reproduces the partition of the space it was derived from, whereas the SCARF embedding deliberately departs from it. This combination — substantially higher cluster separation together with lower agreement with the raw-PV partition — indicates that SCARF does not merely re-express the PV-space regimes more compactly but recovers a finer, better-separated regime structure that the simpler linear pipelines do not capture, justifying the additional cost of contrastive pretraining.

TABLE IV. SIMPLER DIMENSIONALITY-REDUCTION BASELINES VERSUS THE SCARF REPRESENTATION (K-MEANS, K = 15).

Representation	Dim.	Silhouette	ARI vs PV
Standardized process variables (no reduction)	40	0.161	0.474
PCA (95% variance retained)	32	0.174	0.513
PCA (compact)	10	0.236	0.482
SCARF embedding	**128**	**0.366**	**0.354**

[d] All methods cluster the same 162,731-record input (37 standardized process variables plus three mill one-hot indicators) with K-means (k = 15). Silhouette is computed on a fixed 20,000-record subsample; ARI is computed on all records against the k = 8 PV-space reference clustering of Section IV-B.

B. Relationship to PV-Space Regimes

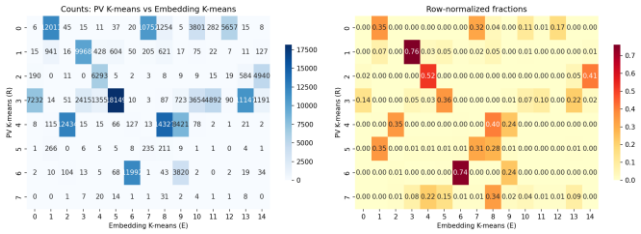


Fig. 5. Cross-tabulation between PV K-means regimes (rows) and embedding K-means clusters (columns). Left: counts; right: row-normalized fractions.

Fig. 5 shows the cross-tabulation between PV-space K-means regimes (R0–R7) and embedding-space K-means clusters (E0–E14), both as raw counts and row-normalized fractions. Two structural patterns are observed. First, certain PV regimes map cleanly to single embedding clusters. PV regime R1 has 76 percent of its samples in embedding cluster E3; PV regime R6 has 74 per cent in E6. These cases correspond to operating regimes that the raw PV space already separates well, and SCARF reinforces this separation. Second, other PV regimes split across multiple embedding clusters. PV regime R0, the largest in the raw PV space, distributes across four embedding clusters (E1, E7, E10, E12) with the largest fraction at 35 percent. PV regime R5 splits across three (E1, E7, E8). This finer subdivision indicates that within what raw PV clustering treats as a single coarse regime, the SCARF embedding identifies sub-regimes that share gross PV characteristics but differ in finer operational structure that the contrastive objective deemed discriminative.

C. Mill Specificity of Embedding Clusters

Fig. 6 displays the mill distribution within each embedding K-means cluster. The pattern is striking: 14 of the 15 clusters are dominated by a single mill (CM2 in clusters 0, 7, 11, 12, 13, 14; CM3 in clusters 1, 3, 5, 10; CM4 in clusters 2, 6, 8, 9). Cluster 4 is the only cluster with appreciable mill mixing, containing approximately 75 percent CM2 and 25 percent CM3 records. Cluster 8 is CM4-dominated (approximately 99 per cent CM4) with a small marginal CM3 component.

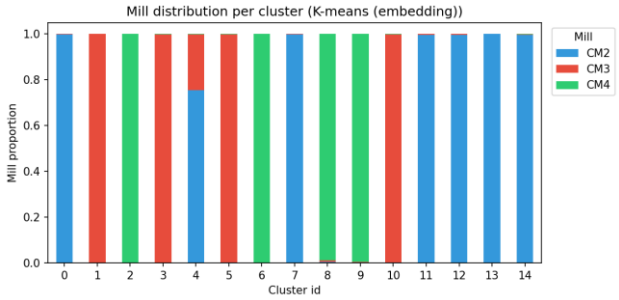


Fig. 6. Mill composition of each embedding K-means cluster. Nearly all clusters are dominated by a single mill, indicating strong mill-specific operating patterns captured by SCARF.

This pattern indicates that SCARF, despite receiving mill identifier only as a 3-dimensional one-hot feature alongside 37 process variables, has learned an embedding in which mill-specific operating patterns dominate. The contrastive objective evidently treats inter-mill differences as larger discriminative signals than within-mill regime differences. Cluster 4, the single mixed cluster, is operationally interpretable: it represents a region of the embedding space where CM2 and CM3 operating modes converge, plausibly indicating a shared operational state achievable on both mills (for example, similar feed composition and similar grinding fineness target).

D. Temporal Evolution of Embedding Clusters

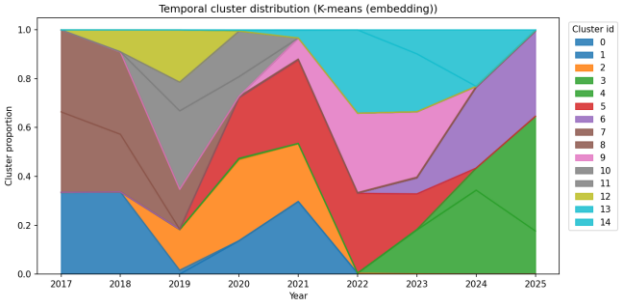


Fig. 7. Year-wise stacked area plot of embedding K-means cluster proportions, showing the progressive evolution of operating modes.

Fig. 7 displays a stacked area plot of cluster proportions by year. The visualization reveals a progressive temporal evolution of operational modes. In 2017 and 2018, clusters 7 and 8 dominate, together accounting for approximately 80 per cent of samples. From 2019 onward, additional clusters (notably clusters 11 and 12) gain prominence and the dominance of

cluster 7 begins to decline. From 2020 to 2021, mid-period clusters (clusters 2 and 5) become prominent. From 2022 onward, a further wave of clusters appears, including clusters 6, 9, 13, and 14. By 2024 and 2025, late-period clusters such as cluster 4 and cluster 6 dominate. The specific color–cluster mapping is shown in the figure legend.

The temporal pattern in Fig. 7 corresponds with the change points reported in Section IV. The first major transition, around 2019, coincides with PELT change points in multiple variables across all three mills. The second major transition, around 2022, coincides with PELT change points in CM2 and CM3. The most recent transition, around 2024, coincides with the latest PELT change point. The combined evidence from PELT (which uses purely temporal information) and SCARF clustering (which uses no temporal information except indirectly through the data distribution) thus converges on the same coarse drift structure, providing independent cross-method validation that the discovered structure reflects genuine operational evolution. We note that the alignment is presented qualitatively (visual coincidence in time) rather than through a formal alignment statistic; a quantitative alignment indicator is discussed as future work in Section VII F.

To quantitatively confirm that the embedding-derived regimes reflect genuine operational differences rather than artifacts of the contrastive learning process, we tested whether the 15 embedding regimes are statistically separable in the original process-variable space. A Kruskal–Wallis test across the regimes was significant for all 37 process variables ($p < 0.001$ in every case), with a mean effect size of $\eta^2 = 0.43$ and values reaching $\eta^2 = 0.79$ for bag-filter differential pressure (MILL_186_BF_DP) and 0.78 for bag-filter pressure (MILL_186_BF_PRESSURE). Because these regimes were discovered purely in the learned embedding space — without access to the raw process-variable distributions during clustering — their strong and uniform separability in process-variable space provides quantitative evidence that the discovered structure corresponds to physically distinct operating modes rather than to learning artifacts.

VII. Discussion

A. Implications for Pre-2023 Data Treatment

The analysis above directly addresses the central methodological question raised in the introduction. The initial restriction to the 2023 to 2025 period was motivated by observations of substantial distributional deviation in pre-2023 Blaine measurements relative to post-2023 records. The present analysis shows that, under a wider valid-range filter applied uniformly across the full 2017 to 2025 period, a meaningful fraction of pre-2023 records does pass both Blaine and 44 μmR range conditions and can be admitted to the labeled set. More importantly, the records that lack valid quality labels (and therefore enter D_unlab) still contain coherent operational structure: PELT detects multiple change points within the pre-2023 period itself (for example, CM3 MILL_SEPOL_SEPOL has six change points distributed across 2017 to 2023), and the SCARF embedding identifies multiple distinct operational clusters within this period (clusters 7, 8, 11, 12 dominate 2017 to 2020). The pre-2023 records thus contribute distinct operational regimes that, if discarded entirely, would represent a substantial loss of information for downstream tasks that benefit from operational diversity, even when those records cannot directly serve as supervised labels.

B. Validity of Embedding-Derived Regimes

The validity of the 15-cluster embedding-derived regime structure is supported by three lines of evidence, each with its own caveat. First, two clustering algorithms with different selection criteria (K-means with silhouette, GMM with BIC) converge on the same number of components ($k = 15$). For K-means, $k = 15$ is an interior optimum of the search grid {3, 5, 8, 10, 12, 15, 20}, which provides direct evidence for $k = 15$. For GMM, $k = 15$ coincides with the upper boundary of its search grid; the agreement is consistent with an intrinsic count but extending the GMM grid would further strengthen this conclusion. Second, the embedding-space silhouette score (0.37) substantially exceeds the raw PV-space score (0.16), although as noted in Section VI A, this comparison is across spaces of different dimensionality and the embedding directly optimizes for contrastive compactness. Third, the temporal evolution of embedding clusters aligns with the PELT change-point locations obtained from a completely independent algorithm using only daily-aggregated time series of three variables. Of the three lines, the third provides the strongest independent corroboration because PELT and SCARF use disjoint information (PELT uses time only on three signals, SCARF uses no time information on 37 signals). The convergence of these assessments provides reasonable, though not definitive, confidence that the discovered regime structure is operationally meaningful.

C. Mill Specificity and Cross-Mill Generalization

The strong mill specificity in Fig. 6 has implications for downstream prediction tasks. If 14 of 15 operating regimes are mill-specific, then models that share parameters across mills must implicitly learn mill-conditional behavior. Cluster 4, the one mixed cluster (CM2/CM3 = 75:25), suggests that some operational states are reachable across mills and that cross-mill transfer is plausible in those regions of the operating space. Future work should investigate whether cross-mill pseudo-labeling within shared operational clusters (such as cluster 4) yields better generalization than facility-wide pseudo-labeling that ignores mill structure.

D. Downstream Quality Prediction Benefit

While the central analysis of this study concerns operational regime structure, we additionally verify that the learned representation provides measurable benefits for the intended downstream task. Using the same labeled set D_l and the SCARF embeddings characterized above, we trained two state-of-the-art deep tabular regressors (FT-Transformer and SAINT) to predict Blaine fineness and 44 μmR, comparing models that receive the raw 40-dimensional input against models that additionally exploit the SCARF representation together with cluster-conditional pseudo-labeling of the unlabeled records. As summarized in Table V, incorporating the learned representation improves test R^2 for every target–model pair, raising Blaine R^2 from 0.524 to 0.567 (FT-Transformer) and from 0.483 to 0.557 (SAINT), and 44 μmR R^2 from 0.836 to 0.852 and from 0.825 to 0.844, respectively. The corresponding reductions in prediction error range from 9.1% to 14.3% and are statistically

significant for all four pairs under both the Wilcoxon signed-rank test and a paired t-test ($p < 0.01$; $n = 7{,}691$). The larger relative gain on Blaine, the harder chemically-mediated target, indicates that the representation is most valuable precisely where raw process variables are least informative. These results confirm that the regime structure recovered through self-supervised pretraining is not merely descriptive but translates into improved predictive performance. A complete end-to-end evaluation of the prediction system — including temporal (chronological) generalization and a systematic study of the pseudo-labeling strategy — is beyond the scope of the present characterization study and is pursued in a separate work.

TABLE V. DOWNSTREAM QUALITY PREDICTION WITH AND WITHOUT SCARF REPRESENTATION

Target	Model	Test R^2 (raw PV)	Test R^2 (+SCARF)	ΔR^2	RMSE (raw→+SCARF)	Wilcoxon p
Blaine	FT-Transformer	0.524	0.567	+0.043	100.7 → 96.0	$<10^{-63}$
Blaine	SAINT	0.483	0.557	+0.074	105.0 → 97.2	$<10^{-89}$
44 µmR	FT-Transformer	0.836	0.852	+0.015	0.521 → 0.496	$<10^{-66}$
44 µmR	SAINT	0.825	0.844	+0.019	0.538 → 0.508	$<10^{-44}$

[e.] Three-seed means: "raw PV" = 40-dimensional input used directly, "+SCARF" = same input with the SCARF representation and cluster-conditional pseudo-labeling.

E. Limitations

Seven limitations should be noted. First, the analysis is conducted on a single facility with three mills; multi-facility validation is essential for assessing generalizability across cement plants with different equipment vendors and operational practices. Within this single facility, however, the three mills (CM2, CM3, and CM4) differ substantially in their operating behavior — as evidenced by the strong mill specificity reported in Section VI C, where 14 of 15 embedding clusters are dominated by a single mill — so the present study already evaluates the method across three distinct operating environments rather than a single homogeneous process. The contribution of this work is methodological: to demonstrate that a coherent operational regime structure can be recovered from multi-year data of a given facility, rather than to claim cross-plant universality of the specific regimes discovered. Validation on independent plants with different equipment vendors remains essential future work to establish broader generalizability. Second, our PELT analysis is restricted to three process variables; a more comprehensive multivariate change-point analysis covering all 37 variables jointly may reveal additional regime transitions not visible in any single variable. Third, the choice of the L2 cost model in PELT assumes mean-shift transitions; transitions characterized by variance changes or covariance shifts may be missed. Fourth, the SCARF contrastive objective treats all feature corruptions as equally informative; feature-specific corruption rates (for example, higher corruption for derived variables and lower for raw sensor readings) may produce more informative embeddings. More broadly, the SCARF hyperparameters used here — the 128-dimensional embedding, the 60 per cent corruption rate, the three-layer encoder, and the contrastive temperature — follow the standard configuration recommended for SCARF rather than values tuned on this dataset; a systematic ablation over these design choices is a natural next step but lies outside the scope of the present characterization study, whose aim is to establish whether a coherent operational regime structure can be recovered rather than to optimize representation-learning hyperparameters. Relatedly, this study evaluates a single self-supervised method (SCARF), which was chosen as a representative contrastive approach well suited to the heterogeneous, mixed-scale process variables of finish milling; a controlled benchmark of SCARF against other self-supervised tabular paradigms — such as the reconstruction-based VIME [9] and SubTab [11], or the attention-based TabTransformer [12] — on the same industrial dataset would help establish which self-supervised inductive bias best captures operational-regime structure and is left to future work. Fifth, while Section VII-D confirms that the learned representation improves downstream quality prediction, a complete end-to-end predictive evaluation — including strictly chronological temporal generalization and a systematic analysis of the pseudo-labeling strategy — is left to separate work. Sixth, the SSL pretraining set described in Section V combines D_1 (including its test split) with D_unlab, which is acceptable for the unsupervised analysis presented in this study but would create a label-free form of test exposure if the same embeddings were reused without modification for downstream supervised evaluation; subsequent supervised studies will retrain the SSL encoder with the supervised test split excluded. Seventh, the regime-stratified random split of D_1 described in Section III-B places temporally adjacent records into different subsets, so any held-out test split constructed from D_1 provides an optimistic estimate of temporal generalization relative to a strictly chronological split; this protocol is adopted in the present study to ensure regime-balanced evaluation across the full 2017 to 2025 data span, and a strictly chronological split is recommended for any downstream supervised study that aims to assess temporal generalization on its own merits.

F. Toward Quantitative Alignment Indicators

The PELT–SCARF alignment reported in Section VI D is presented qualitatively. A quantitative indicator could be constructed by computing, for each PELT change-point date, the local Jensen–Shannon divergence between cluster assignment distributions in fixed-width windows before and after the change point, or by computing the ARI between cluster assignments in different temporal segments. Development of such indicators is left as future work.

VIII. CONCLUSION

This study has characterized the operational drift structure in approximately nine years of cement finish milling DCS data (188,858 hourly records across three industrial mill units) and demonstrated that self-supervised pretraining can recover this structure for downstream applications. PELT change-point detection identifies between 12 and 21 detected drift events per mill, with effect sizes (Cohen's d) consistently satisfying $|d| > 1.0$ for the majority of transitions. SCARF contrastive learning

on the combined labeled and unlabeled set (162,731 records) yields stable 128-dimensional embeddings (validation InfoNCE 6.93 ± 0.02 across three random seeds). Three independent clustering algorithms on the embedding space (K-means, GMM, HDBSCAN) consistently identify 15 to 21 operational regimes with silhouette scores between 0.36 and 0.41, substantially exceeding the raw PV-space silhouette of 0.16 (with the caveats on across-space comparison noted in Section VI-A). The embedding-space K-means regimes exhibit strong mill specificity (14 of 15 clusters dominated by a single mill) and show temporal evolution that aligns with the independently detected PELT change-point boundaries. These findings establish that long-term cement process data contains a richer operational regime structure than implied by raw PV clustering, and that contrastive self-supervised pretraining recovers this structure efficiently from unlabeled records, providing a representation suitable for use in downstream quality prediction tasks (downstream supervised performance is the subject of separate ongoing work).

Two directions for future work follow directly from these results. First, the cluster-conditional pseudo-labeling of the unlabeled subset, using the embedding-derived regimes as conditioning variables, is currently being investigated as a means of enabling state-of-the-art deep tabular models (FT-Transformer, SAINT) to close the performance gap with tree-based ensembles documented in the tabular learning literature; full empirical results will be reported in a forthcoming study. Second, the multi-facility generalization of the discovered regime structure remains an open question; comparison of embeddings across cement facilities with different equipment vendors would clarify whether the SCARF representation captures facility-specific or process-universal regime information.

REFERENCES

[1] A. Hasanbeigi, L. Price, and E. Lin, "Emerging energy-efficiency and CO_2 emission-reduction technologies for cement and concrete production: a technical review," Renewable and Sustainable Energy Reviews, vol. 16, no. 8, pp. 6220–6238, 2012.

[2] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," Journal of the American Statistical Association, vol. 107, no. 500, pp. 1590–1598, 2012.

[3] D. Bahri, H. Jiang, Y. Tay, and D. Metzler, "SCARF: self-supervised contrastive learning using random feature corruption," in Proc. International Conference on Learning Representations (ICLR), 2022.

[4] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: the role of machine learning," IEEE Access, vol. 5, pp. 20590–20616, 2017.

[5] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3235–3243, 2018.

[6] A. K. Pani and H. K. Mohanta, "Soft sensing of particle size in a grinding process: application of support vector regression, fuzzy inference, and adaptive neuro-fuzzy inference techniques for online monitoring of cement fineness," Powder Technology, vol. 264, pp. 484–497, 2014.

[7] X. Yuan, Y. Wang, C. Yang, Z. Ge, Z. Song, and W. Gui, "Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes," IEEE Transactions on Industrial Electronics, vol. 65, no. 2, pp. 1508–1517, 2018.

[8] H. Zermane and A. Drardja, "Development of an efficient cement production monitoring system based on the improved random forest algorithm," International Journal of Advanced Manufacturing Technology, vol. 120, pp. 1853–1866, 2022.

[9] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "VIME: extending the success of self- and semi-supervised learning to tabular domain," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.

[11] T. Ucar, E. Hajiramezanali, and L. Edwards, "SubTab: subsetting features of tabular data for self-supervised representation learning," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021.

[12] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: tabular data modeling using contextual embeddings," arXiv preprint arXiv:2012.06678, 2020.

[13] I. Rubachev, A. Alekberov, Y. Gorishniy, and A. Babenko, "Revisiting pretraining objectives for tabular deep learning," arXiv preprint arXiv:2207.03208, 2022.

[14] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," Signal Processing, vol. 167, p. 107299, 2020.

[15] D. C. Montgomery, Introduction to Statistical Quality Control, 6th ed. Hoboken, NJ: Wiley, 2009.

[16] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: a review," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2346–2363, 2019.

Digital-to-Physical Transfer of Adversarial Patches for Aerial Vehicle Detection

Jung Heum Woo, Eun-Kyu Lee*
College of Information Technology, Incheon National University, Incheon, Republic of Korea

***Abstract*—Deep neural network (DNN)-based object detectors are widely used for analyzing aerial and satellite imagery in applications such as environmental monitoring and urban analytics. Despite their strong performance, these models are known to be vulnerable to adversarial examples, and physical adversarial attacks using printable patterns pose realistic security threats. This study evaluates physical adversarial patch attacks against an aerial vehicle detector by bridging digital optimization and real-world deployment. Adversarial patches are optimized in the digital domain using a loss function that minimizes the maximum objectness score while incorporating non-printability score (NPS) and total variation (TV) constraints to ensure both printability and spatial smoothness. The optimized patches are printed and deployed in three configurations: ON, OFF, and OFF-Side. Experiments using a YOLOv3 detector show that while the OFF patch achieves the highest effectiveness in the digital domain (85.51% Average Objectness Reduction Rate (AORR)), the ON patch demonstrates superior robustness in physical environments (0.197–0.343 Objectness Score Ratio (OSR)) due to its consistent visibility. Furthermore, the results indicate that weather-based augmentation does not necessarily improve patch optimization in this domain. These findings provide critical insights into the practical vulnerabilities of aerial object detection systems.**

***Keywords*—Aerial object detection; physical adversarial attack; adversarial patch; patch optimization; security; attack; feature**

I. Introduction

The rapid growth of aerial and satellite imagery has significantly expanded the use of automated visual analysis systems in a wide range of applications, including environmental monitoring, urban planning, traffic analysis, and disaster response. To efficiently analyze such large-scale image data, DNN-based object detection models have been widely adopted. Modern object detectors have demonstrated strong performance in identifying objects such as vehicles, buildings, and infrastructure in aerial imagery. Among these models, YOLOv3 has been widely used due to its high detection accuracy and real-time inference capability [1].

Despite their strong performance, DNN-based detectors are known to be vulnerable to adversarial examples, which are carefully crafted perturbations designed to manipulate model predictions. Previous studies have shown that adversarial perturbations can significantly degrade the performance of both image classification and object detection models. While early research mainly focused on digital adversarial attacks that directly modify pixel values in images, recent work has demonstrated the feasibility of physical adversarial attacks, in which adversarial patterns are printed and placed in real-world environments to influence camera-captured inputs.

However, most existing studies on physical adversarial attacks have focused on ground-based scenarios, such as autonomous driving systems or surveillance cameras. In contrast, aerial and satellite imagery present unique challenges due to large observation distances, varying viewing angles, scale variations, and environmental conditions such as illumination changes and weather effects. As a result, the effectiveness and transferability of physical adversarial attacks in aerial imagery remain relatively underexplored.

Unlike ground-level photography, aerial and satellite imagery is captured through long atmospheric paths, making it susceptible to visibility degradation caused by weather conditions such as fog, haze, and fluctuating illumination. Such environmental variability can significantly undermine the effectiveness of adversarial patches when they are transferred from digital optimization to physical deployment. To address this, this study incorporates weather-based augmentation into the patch optimization pipeline to simulate atmospheric effects and environmental noise, thereby investigating its impact on ensuring the robustness of adversarial attacks under diverse real-world capture conditions.

Building on this motivation, this study evaluates physical adversarial patch attacks against an aerial vehicle detector by developing and analyzing a practical attack pipeline that bridges digital optimization and real-world deployment. Adversarial patches are first optimized in the digital domain and then physically printed and deployed in real environments to evaluate their impact on object detection performance. The evaluation considers three patch configurations: ON patch placed directly on the vehicle roof, OFF patch placed around the vehicle, and OFF-Side patch in which two patches are placed on both sides of the vehicle. The target detector is pre-trained on the Microsoft COCO dataset [2] and fine-tuned on the Cars Overhead With Context dataset [3] for aerial vehicle detection.

The experimental results yield several critical insights into the physical robustness of aerial vehicle detectors. First, a significant discrepancy is observed between domains: while the OFF patch configuration—which manipulates contextual information—is most effective in digital simulations (achieving up to 85.51% AORR), the ON patch demonstrates superior effectiveness and robustness in real-world environments, reaching a minimum OSR of 0.197. Second, the analysis indicates that weather-based augmentation during the optimization stage does not necessarily improve attack performance and can even hinder the generation of effective adversarial patterns in this specific domain. Finally, moderate total variation (TV) regu-

*Corresponding author
This work was supported by Incheon National University Research Grant in 2026.

larization is shown to be essential for balancing the trade-off between physical printability and adversarial potency. These findings provide insights into the practical vulnerabilities of aerial object detection systems under real-world physical perturbations.

Main contributions of this study are as follows:

- This study presents a digital-to-physical evaluation of adversarial patch attacks for aerial vehicle detection, focusing on the practical transfer gap between digital optimization and real-world aerial capture.
- Object-level and context-level patch deployment strategies, including ON, OFF, and OFF-Side configurations, are compared to show that the most effective strategy differs between digital and physical domains.
- Weather-based augmentation and TV regularization are analyzed in aerial patch optimization, providing empirical evidence that environmental augmentation does not necessarily improve attack effectiveness in this setting.

II. Background and Related Work

A. Aerial Image Object Detection

Object detection in aerial and satellite imagery has become an important research topic due to its wide range of applications, including urban analysis, environmental monitoring, and traffic understanding. Compared with natural-image object detection, aerial image detection is more challenging because target objects often appear very small, densely distributed, and visually ambiguous under varying viewpoints and imaging conditions. To address these challenges, convolutional neural network (CNN)-based detectors have been widely adopted in remote sensing tasks.

Among early large-scale applications, Albert et al. [4] demonstrated that convolutional networks combined with satellite imagery can effectively capture large-scale urban patterns, showing the practical value of deep learning for aerial scene understanding. In addition, Ji et al. [5] proposed a vehicle detection framework for remote sensing imagery that integrates super-resolution with object detection to improve the detection performance of small vehicles in aerial images. These studies demonstrate that deep learning-based approaches have become the dominant paradigm for aerial object detection.

Recent studies have further improved aerial and remote sensing object detection by addressing the unique characteristics of overhead imagery. Cai et al. [6] proposed PKINet to capture multi-scale local features and long-range contextual information for remote sensing detection. Huang et al. [7] investigated detection pre-training for remote sensing scenes and showed that object-level representation alignment can improve transfer performance, especially under limited annotation settings. These advances indicate that aerial object detectors are becoming increasingly specialized for scale variation, dense object layouts, and arbitrary orientations.

More broadly, the success of deep neural networks in aerial imagery has motivated the development of detection systems for vehicles and other small objects observed from overhead viewpoints. In this context, aerial object detection serves as a critical foundation for many downstream tasks, and its reliability is particularly important when the detection results are used in real-world monitoring or decision-making systems.

Despite their strong performance, aerial image detectors remain vulnerable to distribution shifts caused by viewpoint variation, scale change, illumination, and environmental conditions. These limitations become even more critical in adversarial settings, where carefully designed perturbations can exploit the detector's sensitivity. Therefore, understanding the vulnerability of aerial object detectors is necessary not only from the perspective of model accuracy, but also from the viewpoint of security and robustness.

B. Adversarial Attacks on Object Detection

Adversarial examples are intentionally crafted perturbations that cause machine learning models to produce incorrect predictions. Goodfellow et al. [8] established one of the foundational formulations of adversarial examples and showed that deep neural networks are highly vulnerable even to small, human-imperceptible perturbations. Subsequent studies such as DeepFool and Universal Adversarial Perturbations further expanded this line of work by proposing effective optimization-based attack methods and demonstrating that perturbations can generalize across many inputs [9], [10]. Carlini and Wagner [11] later introduced stronger optimization-based attacks that became a standard benchmark for evaluating model robustness.

While many early studies focused on image classification, later work extended adversarial attacks to object detection, which is a more complex task involving both localization and classification. Eykholt et al. [12] examined attacks on object detectors using adversarial stickers, while Song et al. [13] further discussed physical adversarial examples against object detection models. Chen et al. [14] proposed ShapeShifter, a robust physical adversarial attack against Faster R-CNN, showing that detector-specific attacks can remain effective under transformations. Lee and Kolter [15] also investigated physical adversarial patches for object detection, emphasizing the challenges of patch placement and transferability in detection scenarios.

In the aerial domain, den Hollander et al. [16] explored adversarial patch camouflage against aerial detection, suggesting that object detectors operating on overhead imagery are also susceptible to localized perturbations. Similarly, Czaja et al. [17] discussed adversarial examples in remote sensing and highlighted the importance of studying adversarial robustness in aerial and satellite imagery. These studies indicate that adversarial attacks on object detection are not limited to ground-based scenes and that overhead detection systems constitute an important yet relatively underexplored attack surface.

C. Physical Adversarial Patch Attacks

Physical adversarial patch attacks aim to preserve adversarial effectiveness after digital perturbations are transferred into the real world through printing, placement, and image recapture. Kurakin et al. [18] presented one of the earliest demonstrations that adversarial examples can survive the physical imaging process, although their effectiveness may degrade due to viewpoint and environmental variations. To address these

limitations, Brown et al. [19] introduced the adversarial patch paradigm, showing that a localized patch can be optimized to produce universal and robust attack behavior under different transformations. This work became a key foundation for later physical patch attacks.

Subsequent studies improved the realism and robustness of physical adversarial patches by explicitly considering printability, smoothness, and physical-world constraints. Komkov and Petiushko [20] proposed AdvHat, in which physically wearable adversarial patterns were designed to attack face recognition systems. In the object detection domain, many studies extended the patch-based attack setting to detection models, demonstrating that adversarial patterns can suppress or alter detections when placed in physically meaningful regions of the scene [12], [13], [14], [15], [21].

Recent studies have further expanded physical adversarial attacks by considering more realistic deployment factors. Guesmi et al. [22] proposed a dynamic adversarial patch that maintains a naturalistic appearance while improving robustness to transformations such as clothing creases. Wei et al. [23] revisited physical patch attacks from the perspective of camera image signal processing and showed that cross-camera variability can strongly affect attack stability. Long et al. [24] extended printable patch attacks from frame-level detection to multiple object tracking by disrupting both detection and temporal identity association. Liang et al. [25] proposed adversarial camouflage that addresses viewpoint conflicts, illumination changes, and texture smoothness for physical object detection evasion. These studies indicate that physical adversarial attacks are increasingly moving toward realistic imaging, motion, and environmental constraints.

Compared with ground-level physical attacks, physical patch attacks in aerial imagery are more challenging because the attack must remain effective across larger observation distances, severe scale changes, viewpoint differences, and environmental factors such as illumination and weather. Literatures provided early evidence that remote sensing models can be attacked using adversarial perturbations [17], [16]. More recent remote-sensing studies have begun to focus specifically on the digital-to-physical transfer problem. Chen et al. [26] addressed visual inconsistency between digitally optimized patches and their physical appearance by incorporating image harmonization into patch generation, while Li et al. [27] proposed a physical adversarial background patch for aerial object detection that balances attack effectiveness and patch area using Pareto-efficient optimization. Therefore, physical adversarial patch attacks on aerial object detectors represent an important research direction for understanding the real-world security limitations of overhead vision systems.

III. Threat Model

This study considers a threat model in which an attacker attempts to degrade the performance of an aerial vehicle detector using physically deployable adversarial patches. The attacker places adversarial patterns either directly on the target vehicle or around the vehicle so that the detector fails to recognize the object from aerial imagery. The objective of this threat model is to evaluate whether adversarial patches optimized in the digital domain can be effectively transferred to the physical domain and remain effective under real-world aerial capture conditions.

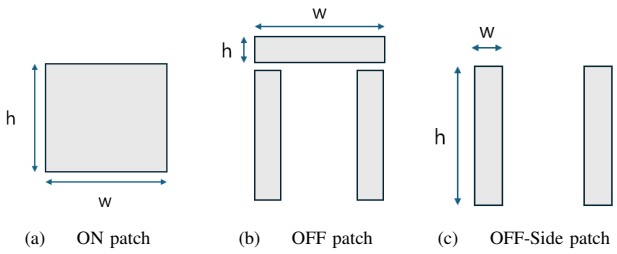


Fig. 1. Illustration of the three patch configurations: (**a**) ON patch, (**b**) OFF patch, and (**c**) OFF-Side patch.

1) Adversarial goal: The primary objective of the attacker is to reduce the detection confidence of an aerial vehicle detector. Specifically, the attacker aims to minimize the objectness score predicted by the detector so that the target vehicle is either not detected or detected with significantly reduced confidence. By decreasing the objectness score, the attack seeks to suppress the detector's confidence in the presence of the vehicle.

2) Knowledge: This study assumes a white-box threat model in which the attacker has full knowledge of the target detection model, including its architecture, parameters, and loss function. Under this assumption, the attacker can optimize adversarial patches directly against the target detector during the digital optimization stage. This setting allows the attack to evaluate the upper bound of adversarial vulnerability under physically realizable patch constraints.

3) Capability: The attacker is capable of generating adversarial patches in the digital domain and deploying them in the physical world. First, the patch is optimized digitally using the target detector. The optimized patch is then printed as a physical pattern and placed in the scene. In this study, three placement strategies are considered: an ON patch placed on the roof of the target vehicle, an OFF patch positioned around the vehicle, and an OFF-Side patch in which two patches are placed on both sides of the vehicle.

4) Strategy: During the optimization, the attacker iteratively updates the adversarial patch so as to minimize the objectness score predicted by the detector. By suppressing the objectness score of the target object, the patch reduces the probability that the detector correctly recognizes the vehicle in aerial imagery. This strategy focuses on degrading detection confidence rather than causing misclassification to another object category.

Under this threat model, the analysis evaluates whether digitally optimized adversarial patches can effectively transfer to the physical domain and degrade the performance of aerial object detectors in real-world capture scenarios.

IV. Attack Method

A. Patch Configuration

To evaluate the effectiveness of adversarial patches under different placement strategies, three patch configurations are

considered in this study, as illustrated in Fig. 1. These configurations represent different physical deployment scenarios that may influence the performance of aerial object detectors.

1) ON patch: The ON patch is placed directly on the roof of the target vehicle. This configuration aims to interfere with the detector by modifying the visual appearance of the object itself. In the digital domain, the ON patch has a size of 200×160 pixels. For physical deployment, the optimized patch is printed at approximately 1189×841 mm so that it covers a substantial portion of the vehicle roof and remains visible from aerial viewpoints.

2) OFF patch: In contrast, the OFF patch is placed around the vehicle rather than directly on it. This configuration attempts to manipulate the contextual information surrounding the target object, thereby affecting the detector's perception of the scene. In the digital domain, the OFF patch is designed as a rectangular strip of size 400×25 pixels. For physical deployment, the patch is printed at approximately 3200×200 mm and arranged around the vehicle to influence the detection process.

3) OFF-Side patch: The OFF-Side patch differs from the OFF patch in both the number and spatial arrangement of the deployed patches. While the OFF patch surrounds the vehicle with three rectangular patches arranged in a Π-shaped structure, the OFF-Side patch places only two patches on the left and right sides of the vehicle. Similar to the OFF configuration, the OFF-Side patch aims to manipulate the contextual region surrounding the vehicle rather than the vehicle surface itself. By perturbing the side regions of the vehicle, this configuration attempts to influence the detector's contextual perception while using fewer patches than the OFF configuration.

These three patch configurations support investigation of how the placement of adversarial patterns affects the robustness of aerial object detectors in both digital simulations and real-world environments. By comparing the ON, OFF, and OFF-Side configurations, this study analyzes the relative effects of object-level perturbations and contextual perturbations on detection performance.

B. Adversarial Patch Optimization

The overall pipeline of the adversarial patch optimization process is illustrated in Fig. 2. To generate adversarial patterns that remain effective in real-world environments, the adversarial patches are first optimized in the digital domain before being deployed in the physical domain. The optimization process aims to reduce the confidence of the object detector by minimizing the objectness score of the target object while ensuring that the generated patch can be reliably printed and physically deployed.

Let P denote the adversarial patch applied to an input image x. During optimization, the patch is iteratively updated to minimize the maximum objectness score produced by the detector. The optimization objective is defined as follows:

$$L_i(P) = \max(\hat{S}_i^u) + \delta \cdot \mathrm{NPS}(P) + \gamma \cdot \mathrm{TV}(P), \qquad (1)$$

where, $\max(\hat{S}_i^u)$ encourages the patch to reduce the highest objectness score in the i-th image, thereby suppressing the detector's confidence for the target object. The second term,

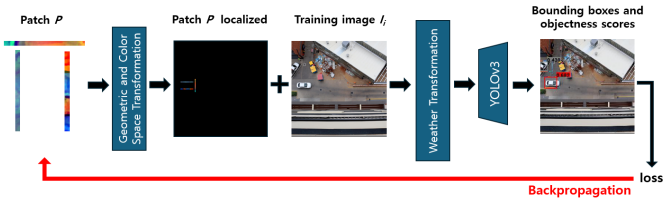


Fig. 2. Overview of the adversarial patch optimization process.

$\mathrm{NPS}(P)$, denotes the Non-Printability Score, which penalizes colors that cannot be reliably reproduced by physical printers. This constraint ensures that the optimized patch remains realizable when printed. The third term, $\mathrm{TV}(P)$, represents total variation regularization, which encourages spatial smoothness by reducing abrupt color changes between neighboring pixels in the patch.

The weighting coefficients δ and γ control the influence of the printability and smoothness constraints, respectively. In the experiments, these parameters are set to $\delta = 0.01$ and $\gamma = 2.5$, following prior work on physical adversarial patches.

In this study, the NPS term is treated primarily as a printability constraint rather than as a main factor for improving adversarial effectiveness. Accordingly, the NPS coefficient is fixed to encourage the optimized patch colors to remain within a physically printable range. In contrast, the TV coefficient is analyzed in more detail because TV regularization directly controls the spatial smoothness of the patch and visibly affects both the optimized patch structure and attack performance. This setting does not fully reveal the interaction between the NPS and TV terms during optimization; therefore, a joint ablation of both coefficients remains an important direction for future work. The default value $\gamma = 2.5$ is selected because, in preliminary experiments, $\gamma = 1.0$ and $\gamma = 2.5$ produced patches that appeared less visually unnatural while maintaining competitive attack performance. Among them, $\gamma = 2.5$ provided a favorable trade-off between visual naturalness and adversarial effectiveness.

To improve robustness against environmental variations, several transformations are applied during patch optimization. These transformations simulate changes that may occur during real-world image capture. Specifically, geometric transformations such as random scaling and rotation are applied to account for viewpoint variations. In addition, color-space transformations, including brightness and contrast adjustments, are introduced to simulate illumination changes. Weather-related perturbations are also considered to approximate atmospheric effects such as lighting variations and environmental noise.

Through this optimization process, the generated adversarial patch is designed to remain effective under both digital evaluation and real-world physical deployment conditions.

V. Detection Model and Dataset

A. Target Detector

This study targets a vehicle detection model designed for aerial imagery. The detector is based on YOLOv3 [1], which is a widely used one-stage object detector known for its efficiency and strong detection performance. YOLOv3 predicts bounding

boxes and objectness scores directly from multi-scale feature maps, making it suitable for detecting small objects such as vehicles in aerial images.

YOLOv3 is selected as the target detector for three main reasons. First, its one-stage detection structure provides efficient inference, which is important for aerial monitoring scenarios where a large number of image frames must be processed. Second, YOLOv3 explicitly predicts objectness scores, which directly align with the objective of this study: suppressing the detector's confidence in the presence of vehicles. This makes the relationship between the optimization loss and the AORR/OSR evaluation metrics transparent and interpretable. Third, although YOLOv3 is not the most recent detector, its explicit objectness-based detection formulation makes it particularly suitable for analyzing adversarial patch optimization and digital-to-physical transferability in the considered aerial vehicle detection setting.

To construct the target detector, the model is first pre-trained on the Microsoft COCO dataset [2], which contains diverse object categories and large-scale visual data. This pre-training stage enables the detector to learn general visual features for object recognition. The detector is then fine-tuned for aerial vehicle detection using the Cars Overhead With Context (COWC) dataset [3]. After fine-tuning, the trained detector is used as the target model for evaluating adversarial patch attacks in both the digital and physical domains.

B. Training Dataset

The target detector is trained using a two-stage dataset configuration consisting of a general-purpose pre-training dataset and an aerial-domain fine-tuning dataset. First, the detector is pre-trained on the Microsoft COCO dataset [2], which is a large-scale benchmark for object detection containing a wide variety of object classes and scene types. This stage provides the detector with general feature representations that are useful for downstream detection tasks.

After pre-training, the detector is fine-tuned on the COWC dataset [3], which is specifically designed for vehicle detection in overhead imagery. The COWC dataset contains more than 25,000 color image patches captured from aerial viewpoints in diverse urban environments. Each image has a spatial resolution of 256×256 pixels and includes vehicle annotations for detection tasks. Since the dataset focuses on overhead vehicle imagery, it is well suited for adapting the detector to the aerial detection setting considered in this study.

Examples from the training data are shown in Fig. 3, including representative scenes from the COWC dataset such as Selwyn and Utah. These images illustrate the overhead viewpoint, small object scale, and scene complexity that characterize aerial vehicle detection.

C. Physical Evaluation Dataset

To evaluate the real-world effectiveness of adversarial patches, additional aerial images are collected in two physical scenarios: *Side Street* and *Car Park*. These datasets are used exclusively for physical-domain evaluation and are intended to assess whether digitally optimized adversarial patches remain effective after physical deployment.

(a) COWC–Selwyn (b) COWC–Utah

(c) Car Park (d) Side Street

Fig. 3. Example images from the datasets used in this study: (**a**) COWC–Selwyn, (**b**) COWC–Utah, (**c**) Car park, and (**d**) Side street.

In the Side Street scenario, aerial images are captured from approximately the height of a ten-story building, corresponding to an observation height of about 40 m. In the Car Park scenario, aerial images are collected using an unmanned aerial vehicle (UAV) flying at an altitude of approximately 60 m. These two settings provide different viewpoints and scene structures, allowing the robustness of the adversarial patches to be evaluated under varying environmental conditions.

During physical data collection, three controlled vehicles are used, namely gray, white, and blue vehicles, depending on the scenario. The images and videos are recorded under multiple environmental conditions, including direct sunlight and shade, as well as static and moving vehicle states. The resulting physical evaluation dataset consists of 5,225 frames for the Side Street scenario and 2,126 frames for the Car Park scenario. Example images from these physical evaluation environments are also shown in Fig. 3.

These datasets enable a comprehensive evaluation of adversarial patch transferability from digital optimization to physical deployment under realistic aerial capture conditions.

VI. Experimental Setup

All experiments in this study were conducted on a workstation equipped with an NVIDIA GeForce RTX 4070 GPU. The models were implemented and trained using PyTorch 2.5.1 with CUDA 12.1.

A. Attack Settings

To evaluate the effectiveness of adversarial patches against aerial vehicle detectors, the patches are first optimized in the

digital domain and then deployed in the physical domain. During optimization, each patch is iteratively updated to minimize the objectness score predicted by the target detector while satisfying the printability and smoothness constraints described in Eq. (1). The optimization objective combines the maximum objectness term with the Non-Printability Score (NPS) and Total Variation (TV) regularization in order to generate adversarial patterns that are both physically realizable and robust under real-world capture conditions.

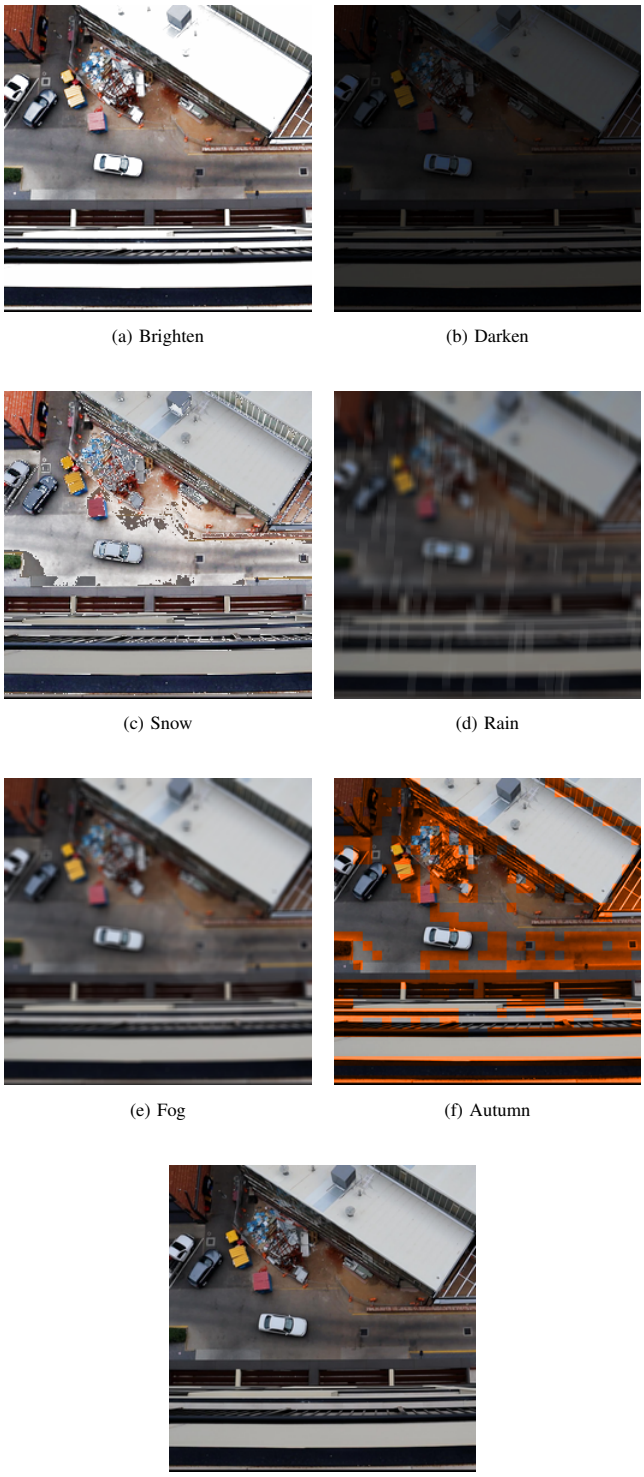
During patch training, several transformations are applied to improve robustness against environmental variations. Specifically, geometric transformations such as random scaling and rotation are used to simulate viewpoint changes in aerial imagery. In addition, color-space transformations including brightness and contrast adjustments are introduced to account for illumination variations. Weather-based augmentation is also incorporated to approximate environmental disturbances such as atmospheric effects and lighting variation.

To analyze the effect of augmentation strategies during patch optimization, three training pipelines are considered in this study: *G/C+W*, *G/C*, and *Control*. The *G/C+W* setting includes geometric transformations, color-space transformations, and weather augmentation. The *G/C* setting includes only geometric and color-space transformations. The *Control* setting uses a random patch without adversarial optimization. In addition, three patch placement strategies are evaluated, namely the ON, OFF, and OFF-Side configurations described in the previous section. These settings allow us to compare how augmentation strategies and patch placement affect attack performance in both digital and physical environments.

B. Weather Augmentation

To evaluate the robustness of adversarial patches under environmental appearance changes, image-level weather augmentation is applied during both patch training and digital testing when weather augmentation is enabled. In this process, the adversarial patch is first geometrically transformed according to the target object location and pasted onto the clean image. Then, the weather transformation is applied to the entire patched image before resizing it to the YOLOv3 input resolution. Therefore, the weather effect influences both the background scene and the adversarial patch, which better approximates physical deployment conditions where illumination, rain, fog, and color shifts affect the whole captured image.

For each image, one weather condition is randomly selected with uniform probability from seven candidates: brightening, darkening, snow, rain, fog, autumn color shift, and identity transformation. This selection corresponds to sampling an integer value from 0 to 6, where the seven cases denote brightening, darkening, snow, rain, fog, autumn color shift, and no transformation, respectively. The same weather augmentation policy is used during digital testing under the *STD+W* setting. Representative examples of the seven weather augmentation cases are shown in Fig. 4.



(a) Brighten (b) Darken

(c) Snow (d) Rain

(e) Fog (f) Autumn

(g) Identity

Fig. 4. Examples of the seven image-level weather augmentation cases used in patch training and digital testing: (**a**) brighten, (**b**) darken, (**c**) snow, (**d**) rain, (**e**) fog, (**f**) autumn color shift, and (**g**) identity transformation.

The brightening and darkening transformations are performed in the HLS color space by modifying the lightness channel. For brightening, the lightness coefficient is sampled as $1+\mathcal{U}(0,1)$, resulting in a scaling factor in the range $[1.0, 2.0]$. For darkening, the coefficient is sampled as $1-\mathcal{U}(0,1)$, producing a scaling factor in the range $[0.0, 1.0]$. The snow transformation increases the lightness of relatively dark pixels. Its threshold is sampled as $\mathcal{U}(0,1) \times 255/2 + 255/3$, corresponding approximately to the range $[85, 212]$, and pixels

below this threshold are brightened using a fixed coefficient of 2.5.

The rain transformation generates random rain streaks over the image. The slant of the streaks is selected from the integer range $[-10, 10]$, while the default drop length, drop width, and drop color are set to 20, 1, and $(200, 200, 200)$, respectively. The number of rain drops is proportional to the image area and is computed as the image area divided by 600. After drawing the rain streaks, a blur operation with a 7×7 kernel is applied, and the lightness channel is multiplied by 0.7 to simulate darker rainy conditions. The fog transformation overlays semi-transparent white circular haze patterns, with the fog coefficient sampled from $\mathcal{U}(0.3, 1.0)$. The haze size is determined as one third of the image width multiplied by the fog coefficient, and the final image is blurred using a kernel size proportional to the haze size. Finally, the autumn transformation divides the image into local 8×8 regions and changes the hue of selected low-green regions to one of four autumn-like hue values, $\{1, 5, 9, 11\}$, thereby simulating seasonal background color shifts.

Overall, the processing order for weather-based experiments is as follows: the clean image is padded to a square shape, the transformed adversarial patch is pasted, weather augmentation is applied to the patched image, the image is resized to 256×256, and YOLOv3 inference is then performed. For non-weather experiments, the same procedure is used except that the weather augmentation step is skipped.

C. Evaluation Metrics

To quantitatively evaluate the effectiveness of adversarial patch attacks, this study adopts objectness-based metrics for both digital-domain and physical-domain experiments. In the digital domain, the attack performance is measured using the Average Objectness Reduction Rate (AORR), which quantifies the relative decrease in objectness score after applying the adversarial patch. AORR is defined as

$$\mathrm{AORR}(T^v, \hat{T}^v) \triangleq \frac{1}{NE_j} \sum_{j=1}^{N} \sum_{l=1}^{E_j} \frac{s^v_{j,l} - \hat{s}^v_{j,l}}{s^v_{j,l}}, \tag{2}$$

where, T^v denotes the set of detection results for all images in the test set v, N is the number of test images, and E_j represents the number of detected objects in the j-th image. The term $s^v_{j,l}$ denotes the original objectness score of the l-th detection in the j-th image, whereas $\hat{s}^v_{j,l}$ denotes the corresponding objectness score after applying the adversarial patch. A higher AORR value indicates a stronger attack effect because it reflects a larger reduction in the detector's confidence.

AORR is used because the adversarial patch is optimized to directly suppress the objectness score of the target detector. Therefore, AORR measures the relative change in the same confidence signal that is minimized during patch optimization and provides a direct view of whether the learned patch successfully weakens the detector's vehicle-related response. Standard detection metrics such as mAP remain important for assessing final detection performance, but they can be affected by thresholding, non-maximum suppression, and localization changes. In contrast, AORR captures confidence degradation before and after patch application at the objectness score level. For this reason, AORR is reported as the primary attack-oriented metric, and mAP is additionally reported in the digital-domain evaluation to verify whether objectness suppression also leads to actual detection-performance degradation.

For physical-domain evaluation, the Objectness Score Ratio (OSR) is used, which measures the relative reduction in detection confidence after deploying the printed adversarial patch. OSR is defined as

$$\mathrm{OSR}(\hat{S}, S) \triangleq \frac{\frac{1}{\alpha} \sum_{f=1}^{\alpha} \hat{S}_f}{\frac{1}{\beta} \sum_{g=1}^{\beta} S_g}, \tag{3}$$

where, S_g denotes the original objectness score of the target object, and $\hat{S}_f$ represents the objectness score after the adversarial patch is applied. The parameter α denotes the total number of frames captured when the adversarial patch is physically deployed, whereas β denotes the total number of frames captured without the adversarial patch. A lower OSR value indicates a more effective physical attack.

OSR is used in the physical-domain evaluation because perfectly paired clean and patched images are difficult to obtain in real aerial capture. Unlike the digital setting, where the same image can be evaluated before and after inserting a patch, physical experiments are affected by small changes in viewpoint, illumination, vehicle position, camera motion, and environmental conditions. Therefore, directly computing an image-wise objectness reduction rate analogous to AORR is not reliable in the physical domain. Instead, OSR compares the average objectness score of frames captured with the printed patch to that of frames captured without the patch, providing a practical measure of confidence reduction under real deployment conditions.

D. Digital-Domain Evaluation

In the digital-domain experiments, adversarial patches are directly inserted into test images to evaluate their effectiveness without physical deployment. The experiments are conducted using the aerial vehicle detector described in the previous section, which is based on the YOLOv3 architecture.

To evaluate the robustness of the optimized patches under different testing conditions, two testing regimes are considered. The first regime, denoted as *STD*, measures AORR on the original test dataset without additional perturbation. The second regime, denoted as *STD+W*, applies weather effects to the test images before evaluation. These two settings allow us to analyze whether the optimized patches remain effective when environmental variations are introduced during testing.

The digital-domain evaluation is performed on the Side Street dataset. The same digital analysis is not conducted on the Car Park dataset because vehicles in that environment are parked very close to one another, causing visual interference among multiple patches and making it difficult to isolate the effect of each patch configuration. For each training pipeline and patch configuration, the attack performance is measured using AORR under both the *STD* and *STD+W* settings.

E. Physical-Domain Evaluation

To evaluate real-world attack effectiveness, this study uses the physical deployment settings provided by the benchmark in [28]. The physical experiments are conducted in two scenarios: *Side Street* and *Car Park*. In the Side Street scenario, images are captured from approximately 40 m above the ground, corresponding to the height of a ten-story building. In the Car Park scenario, aerial images are collected using a UAV flying at an altitude of approximately 60 m.

In this benchmark, the adversarial patches are printed on 160 gsm coated paper and placed either on the roof of the vehicle (ON patch) or around the vehicle (OFF patch). The OFF-Side configuration is considered in digital-domain analysis, whereas the physical-domain experiments mainly focus on the printed ON and OFF configurations. During data collection, the vehicles are recorded under varying environmental conditions, including direct sunlight and shade, as well as static and moving states.

The captured aerial images and video frames are processed by the target detector to obtain objectness scores for the target vehicle. The effectiveness of the adversarial attack is then measured using the OSR metric defined in Eq. (3). Through this evaluation protocol, the analysis assesses whether digitally optimized adversarial patches can maintain their effectiveness after physical printing, placement, and aerial image capture.

VII. Results and Analysis

A. Digital-Domain Results

Digital-domain evaluation reports results only on the Side Street dataset. Analyses on the Car Park dataset are omitted because vehicles in that environment are parked very close to one another, causing multiple patches to visually interfere with each other. This inter-patch interference makes it difficult to isolate the effect of each patch configuration and may distort the evaluation of attack effectiveness.

Table I summarizes the AORR results for the three patch configurations under different training pipelines. Overall, the *G/C* pipeline consistently outperforms the *G/C+W* pipeline for all patch types under both the *STD* and *STD+W* evaluation settings. This result suggests that incorporating weather-based augmentation does not improve adversarial patch optimization in this setting and may even degrade attack effectiveness. Out of the three patch configurations, the OFF patch achieves the strongest overall attack performance in the digital domain. In particular, the OFF patch trained with the *G/C* pipeline attains the highest AORR values, reaching 85.51% under *STD* and 78.02% under *STD+W*. These results indicate that perturbing the contextual region around the vehicle is more effective in the digital domain than directly modifying the appearance of the vehicle itself. The OFF-Side patch also exhibits meaningful attack performance, but it does not surpass the OFF patch. Specifically, the OFF-Side patch achieves 67.43% under *STD* and 62.43% under *STD+W* with the *G/C* pipeline, which is higher than the ON patch in some cases but still lower than the OFF patch overall. These results suggest that attack effectiveness varies depending on the number and spatial arrangement of the deployed patches.

Table I also shows that applying adversarial patches substantially decreases the detection accuracy compared with the clean baseline. Under the *STD* setting, the clean image achieves 56.22% mAP, whereas the adversarially optimized patches reduce the mAP to 5.98%, 7.46%, and 10.85% for the ON, OFF, and OFF-Side configurations with the *G/C* pipeline, respectively. A similar degradation is observed under the *STD+W* setting, where the clean mAP is 34.15%, but the adversarial patch cases remain much lower, ranging from 6.38% to 13.50% depending on the patch configuration and training pipeline. These mAP reductions are consistent with the AORR results: patch configurations with large objectness reduction generally produce severe detection-performance degradation. For example, the OFF patch with the *G/C* pipeline achieves the highest AORR under both *STD* and *STD+W*, while also reducing mAP far below the clean baseline. This consistency indicates that the AORR metric reflects not only confidence suppression but also practical degradation in detection performance. In contrast, the control patches generally retain higher mAP values or exhibit lower AORR values than the adversarially optimized patches, indicating that the performance drop is mainly caused by the optimized adversarial patterns rather than by simple occlusion or patch placement alone.

Overall, the digital-domain results reveal three key findings: 1) weather-based augmentation is not beneficial for patch optimization in this setting, 2) the OFF patch is the most effective configuration in the digital domain, and 3) attack performance changes according to the number and spatial extent of the deployed patches.

B. Physical-Domain Results

The physical-domain results in Table II are based on the aerial patch deployment benchmark reported in [28]. In this benchmark setting, the optimized adversarial patches are printed on 160 gsm coated paper, and aerial videos are recorded at 25 FPS for both the Car Park and Side Street scenarios. During data collection, three controlled vehicles, namely gray, white, and blue vehicles, are equipped with the printed adversarial patches. The resulting dataset consists of 2,126 frames for the Car Park scenario and 5,225 frames for the Side Street scenario.

Table II summarizes the physical-domain attack results. Overall, the deployed patches significantly reduce the detector's objectness scores across different vehicles and environmental conditions. The measured OSR values range from 0.197 to 0.746, indicating substantial degradation in detection confidence after physical patch deployment. Unlike the digital-domain results, where the OFF patch achieves the best performance, the ON patch generally shows stronger effectiveness in the physical domain, with the lowest mean OSR of 0.197 in the Side Street scenario. In the Side Street scenario, the ON patch reduces the mean OSR to values between 0.197 and 0.343 depending on the vehicle, lighting condition, and motion state. In contrast, the OFF patch shows weaker performance in the same scenario, with OSR values of 0.429 and 0.748 in the evaluated settings. In the Car Park scenario, the ON patch also reduces the detector confidence, achieving OSR values of 0.208, 0.509, and 0.746 for different vehicles and lighting conditions.

TABLE I. DIGITAL-DOMAIN EVALUATION RESULTS OF ADVERSARIAL PATCHES ON THE SIDE STREET DATASET UNDER BOTH STD AND STD+W EVALUATION SETTINGS. FOR PATCHED CASES, BOLD AND UNDERLINE INDICATE THE BEST AND SECOND-BEST ATTACK RESULTS, RESPECTIVELY; LOWER MAP AND HIGHER AORR INDICATE STRONGER ATTACKS

Training pipeline	Patch type	mAP (STD) [%]	AORR (STD) [%]	mAP (STD+W) [%]	AORR (STD+W) [%]
–	Clean	56.22	–	34.15	–
G/C	ON	5.98	72.59	7.07	64.09
G/C+W	ON	**4.10**	53.76	**6.38**	54.44
Control	ON	20.01	29.06	16.48	34.75
G/C	OFF	7.46	**85.51**	9.52	**78.02**
G/C+W	OFF	<u>4.93</u>	<u>79.37</u>	<u>6.57</u>	<u>71.80</u>
Control	OFF	27.28	20.65	18.01	10.73
G/C	OFF-Side	10.85	67.43	13.50	62.43
G/C+W	OFF-Side	5.54	66.46	12.28	59.36
Control	OFF-Side	39.65	15.66	28.86	28.49

TABLE II. RESULTS OF ADVERSARIAL PATCHES IN THE PHYSICAL DOMAIN [28]

(a) Side Street

Car	Patch type	Lighting	Motion	Mean OSR
Gray	ON	Both	Moving	0.343
Gray	ON	Sun	Static	0.251
Gray	ON	Shade	Static	0.255
Gray	OFF	Sun	Static	0.429
White	ON	Both	Moving	0.286
White	ON	Sun	Static	0.285
White	ON	Shade	Static	0.197
White	OFF	Sun	Static	0.748

(b) Car Park

Car	Patch type	Lighting	Motion	Mean OSR
Gray	ON	Sun	Static	0.509
Blue	ON	Shade	Static	0.208
White	ON	Sun	Static	0.746

These results show that adversarial patches optimized in the digital domain can be transferred to real-world aerial environments and remain effective after physical deployment. At the same time, the results indicate that the most effective patch configuration differs between the digital and physical domains.

C. Effect of TV Regularization

To analyze the effect of the TV regularization term, the coefficient γ in the patch optimization loss is varied, and the OFF patch is evaluated under the *G/C+W* training pipeline. The results are summarized in Table III, and representative visual examples of the optimized patches are shown in Fig. 5.

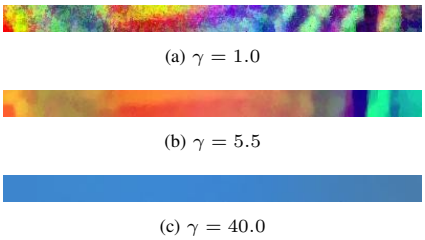
Overall, the attack performance strongly depends on the value of the TV loss coefficient. When $\gamma = 0.0$, the AORR is relatively low, reaching 33.18% under *STD* and 36.76% under *STD+W*, indicating that patches without smoothness regularization are less effective. Introducing TV regularization significantly improves the attack performance, with the best results observed around $\gamma = 1.0$ and $\gamma = 2.5$. Specifically, the highest AORR values are obtained at $\gamma = 1.0$ under *STD* and at $\gamma = 2.5$ under *STD+W*.

TABLE III. RESULTS OF ADVERSARIAL PATCHES UNDER THE G/C+W TRAINING PIPELINE WITH DIFFERENT TV LOSS COEFFICIENTS. BOLD AND UNDERLINE INDICATE THE BEST AND SECOND-BEST ATTACK RESULTS, RESPECTIVELY

TV loss coefficient (γ)	AORR (STD) [%]	AORR (STD+W) [%]
0.0	33.18	36.76
1.0	**80.79**	<u>71.32</u>
2.5	<u>79.37</u>	**71.80**
4.0	76.23	69.17
5.5	72.54	66.72
20.0	53.52	56.26
40.0	47.19	53.23

However, as the value of γ increases further, the attack performance gradually decreases. For example, when $\gamma = 20.0$ or $\gamma = 40.0$, the AORR drops substantially in both evaluation settings. This trend indicates that excessive TV regularization overly smooths the patch, thereby weakening the adversarial patterns that are necessary for suppressing the detector's objectness score.

The visual examples in Fig. 5 support this observation. When γ is small, the optimized patch contains diverse and high-frequency color patterns that effectively disrupt the detector. As γ increases, the patch becomes progressively smoother, and at very large values, such as $\gamma = 40.0$, the patch converges to a nearly uniform color pattern. These results indicate that moderate TV regularization is necessary to balance printability and attack effectiveness, whereas excessive smoothing degrades adversarial performance.



(a) $\gamma = 1.0$

(b) $\gamma = 5.5$

(c) $\gamma = 40.0$

Fig. 5. Visual examples of optimized patches under different TV loss coefficients: (**a**) $\gamma = 1.0$, (**b**) $\gamma = 5.5$, and (**c**) $\gamma = 40.0$.

VIII. Discussion

This study focuses on evaluating the digital-to-physical transfer of adversarial patches in aerial vehicle detection, where attacks are affected by overhead viewpoints, small object scale, long-distance capture, physical visibility, and context dependence. These factors distinguish the aerial setting from classic ground-based physical patch scenarios.

The experimental results provide several important observations regarding the transferability of adversarial patches from digital optimization to physical deployment in aerial vehicle detection. First, the digital-domain results show that the OFF patch achieves the strongest attack performance, whereas the physical-domain results indicate that the ON patch is generally more effective after real-world deployment. This discrepancy suggests that the attack effectiveness in the digital domain does not necessarily translate directly to the physical domain. One possible explanation for this difference is the distinct role of contextual perturbations and object-level perturbations under real-world imaging conditions. In the digital domain, the OFF patch is highly effective because it perturbs contextual cues around the vehicle, such as road texture, parking-lot structure, shadows, and local background contrast, while its position, scale, and visibility are precisely controlled. However, this context-dependent attack becomes less stable after physical deployment, where viewpoint changes, partial occlusion, scale variation, boundary misalignment, environmental noise, and small placement errors can alter the spatial relationship between the vehicle and the surrounding patch. Since the OFF patch is placed around the vehicle, nearby objects, occlusions, and shadows may further reduce its visibility and explain why the ON patch appears more effective in the physical setting. In contrast, the ON patch is attached directly to the vehicle body and moves together with the target object, maintaining more consistent visibility and object-level alignment across aerial frames. This explains why the ON patch is more robust in the physical domain, even though the OFF patch is stronger in digital evaluation. The OFF-Side configuration further shows the role of spatial coverage, as perturbing only the two side regions provides weaker contextual manipulation than the full OFF configuration. Since OFF-Side is mainly evaluated digitally in this study, a complete physical comparison among ON, OFF, and OFF-Side remains future work.

Second, the results show that weather-based augmentation during optimization does not improve attack performance and may even reduce it. In the digital-domain experiments, the *G/C* pipeline consistently outperforms the *G/C+W* pipeline for all patch configurations. This finding suggests that the additional variability introduced by weather augmentation might hinder the optimization of discriminative adversarial patterns in the considered setting. Rather than improving robustness, excessive augmentation may weaken the attack objective by forcing the patch to adapt to overly broad or unrealistic perturbation conditions. A possible mechanism behind this result is that the weather transformation perturbs both the adversarial patch and the surrounding scene after patch placement. Because the attack objective is based on suppressing objectness scores, the optimizer must learn patch patterns that consistently disturb vehicle-related features across many transformed appearances. Strong image-level weather effects such as fog, rain, snow, and large brightness changes can alter edge contrast, local texture statistics, and color distributions of the entire image. As a result, gradients obtained from different weather conditions may become less consistent across iterations, making the optimization less stable and reducing the emergence of sharp, discriminative adversarial patterns. In addition, weather augmentation may partially mask or smooth the patch itself, especially when haze, rain blur, or strong illumination changes are applied after patch insertion. This can weaken the feature perturbation produced by the patch and encourage the optimization to converge to patterns that are more generic but less effective against the detector's stable vehicle features. Therefore, in this setting, weather augmentation appears to increase appearance diversity but does not necessarily improve adversarial generalization. Instead, it can dilute the attack signal and reduce optimization efficiency. Therefore, the effect of weather augmentation should be interpreted as an empirical observation in this setting, and further analysis of convergence behavior, gradient stability, or detector feature responses would be needed to explain the underlying optimization mechanism.

Third, the ablation study on TV regularization demonstrates that moderate smoothness constraints are essential for effective physical adversarial patch generation. When the TV coefficient is too small, the optimized patch lacks sufficient structural regularity for stable physical deployment. On the other hand, when the coefficient becomes too large, the patch becomes overly smooth and loses the high-frequency patterns necessary for suppressing detector confidence. This result indicates that there is a trade-off between printability and adversarial effectiveness, and that an intermediate level of regularization provides the best balance.

Overall, these findings highlight several practical implications for aerial object detection security. They show that aerial detectors are vulnerable not only to digitally optimized perturbations but also to physically deployed patches that remain effective under real-world capture conditions. At the same time, the results reveal that the most effective attack strategy depends strongly on whether the evaluation is conducted in the digital or physical domain. Therefore, robustness assessment for aerial object detectors should not rely solely on digital experiments, but should also consider physical deployment scenarios and environmental variability.

Despite these findings, this study has limitations. First, the experiments are conducted using a single detector architecture, YOLOv3, and the generality of the results to other aerial object detectors remains to be verified. In particular, whether the observed digital-to-physical transfer behavior extends to anchor-free detectors or transformer-based detectors is not investigated in this study. Second, the physical evaluation is limited to a small number of vehicles and capture environments. Third, the OFF-Side configuration is mainly evaluated in the digital domain, and further physical experiments would provide a more complete comparison among all patch types. In addition, the optimization hyperparameters are fixed after preliminary experiments, and a full sensitivity analysis over learning rate, initialization strategy, and optimization duration is not conducted. The reported AORR and OSR values are descriptive averages, and repeated optimization runs, confidence intervals, standard deviations, and significance tests are not included. In addition, the current weather analysis does not explicitly consider dynamic shadows or partial patch occlusion caused

by buildings, trees, or changes in sun position, although such effects are important in real aerial photography. Future work should extend the analysis to additional detector architectures, more diverse aerial scenarios, dynamic shadow and partial-overlap conditions, and defense strategies for improving robustness against physical adversarial perturbations.

IX. Conclusion

This study investigated the transferability of adversarial patches from digital optimization to physical deployment for aerial vehicle detection. Adversarial patches were optimized in the digital domain using an objectness-based loss with Non-Printability Score and Total Variation regularization, and were then evaluated under three placement strategies: ON, OFF, and OFF-Side.

The experimental results showed that the OFF patch achieved the strongest attack effectiveness in the digital domain, indicating that contextual perturbations around the vehicle can strongly suppress detector confidence in simulation-based evaluation. In contrast, the ON patch was more effective in the physical domain, suggesting that object-level perturbations are more robust after printing and real-world deployment. In addition, weather-based augmentation did not improve patch optimization performance in this setting, and the TV loss coefficient was found to have a substantial influence on attack effectiveness.

These results demonstrate that digitally optimized adversarial patches can transfer to real-world aerial environments and significantly degrade the performance of aerial vehicle detectors. More importantly, the study reveals that the most effective patch configuration differs between digital and physical evaluations. This finding highlights the importance of considering physical deployment conditions when assessing the robustness of aerial object detection systems.

Overall, this work provides empirical evidence of the practical vulnerability of aerial object detectors to physical adversarial attacks and offers useful insights for future research on robust aerial perception and adversarial defense.

References

[1] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[3] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 785–800.

[4] A. Albert, J. Kaur, and M. C. Gonzalez, "Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[5] H. Ji, Z. Gao, T. Mei, and B. Ramesh, "Vehicle detection in remote sensing images leveraging on simultaneous super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, pp. 676–680, 2019.

[6] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao, "Poly kernel inception network for remote sensing detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 706–27 716.

[7] Z. Huang, Y. Feng, Q. Liu, and Y. Wang, "Mutdet: Mutually optimizing pre-training for remote sensing object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 1–17.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[9] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] ——, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2017.

[12] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, D. Song, T. Kohno, A. Rahmati, A. Prakash, and F. Tramer, "Note on attacking object detectors with adversarial stickers," *arXiv preprint arXiv:1712.08062*, 2017.

[13] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno, "Physical adversarial examples for object detectors," in *Proceedings of the 12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018.

[14] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2018.

[15] M. Lee and Z. Kolter, "On physical adversarial patches for object detection," in *Proceedings of the ICML Workshop on Security and Privacy of Machine Learning*, 2019.

[16] R. Den Hollander, A. Adhikari, I. Tolios, M. van Bekkum, A. Bal, S. Hendriks, M. Kruithof, D. Gross, N. Jansen, G. Perez *et al.*, "Adversarial patch camouflage against aerial detection," in *Artificial Intelligence and Machine Learning in Defense Applications II*, vol. 11543. SPIE, 2020.

[17] W. Czaja, N. Fendley, M. Pekala, C. Ratto, and I.-J. Wang, "Adversarial examples in remote sensing," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018.

[18] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[19] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[20] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, 2021.

[21] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[22] A. Guesmi, R. Ding, M. A. Hanif, I. Alouani, and M. Shafique, "Dap: A dynamic adversarial patch for evading person detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 595–24 604.

[23] H. Wei, Z. Wang, K. Zhang, J. Hou, Y. Liu, H. Tang, and Z. Wang, "Revisiting adversarial patches for designing camera-agnostic attacks against person detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 8047–8064, 2024.

[24] J. Long, T. Jiang, W. Yao, S. Jia, W. Zhang, W. Zhou, C. Ma, and X. Chen, "Papmot: Exploring adversarial patch attack against multiple object tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 128–144.

[25] J. Liang, S. Liang, T. Lou, M. Zhang, W. Li, D. Fan, and X. Cao, "Gradient-reweighted adversarial camouflage for physical object detection evasion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 13 880–13 889.

[26] J. Chen, Y. Zhang, C. Liu, K. Chen, Z. Zou, and Z. Shi, "Digital-to-physical visual consistency optimization for adversarial patch generation in remote sensing scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[27] H. Li, J. Li, M. Gong, H. Yu, K. Dang, Y. Zhou, A. K. Qin, and Y. Wu, "Physical adversarial background patch against aerial object detection based on pareto efficiency," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[28] A. Du, B. Chen, T.-J. Chin, Y. W. Law, M. Sasdelli, R. Rajasegaran, and D. Campbell, "Physical adversarial attacks on an aerial imagery object detector," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

A Hybrid Ethereum-Based Architecture for Secure Electronic Health Records: Consent, Integrity Anchoring and Auditable Access

Rodica Doina Zmaranda, Attila-Imre Kovacs, Daniela Elena Popescu, Alexandrina Mirela Pater
Department of Computer Science and Information Technology, University of Oradea, Oradea, Romania

***Abstract*—Securing electronic health records (EHR) requires strong guarantees for confidentiality, integrity, access control, and auditability. Traditional centralized architectures rely on database-level protection and internal logging, which remain vulnerable to insider misuse and undetected data modification. This study proposes a practical hybrid architecture in which medical content is stored encrypted off-chain, while blockchain is used selectively as a governance and evidence layer. An Ethereum-based prototype was designed and implemented to support integrity anchoring of medical documents, patient-controlled consent management, and immutable audit trails for critical actions. In the implemented solution, the actual medical content is not stored on-chain. Instead, the blockchain stores only document-related metadata, cryptographic hashes, document references, and access-control information, while the sensitive medical data remains encrypted and stored off-chain. This design supports GDPR-oriented data minimization, since the immutable blockchain layer does not contain raw medical records or directly identifiable medical content. The prototype separates confidentiality from blockchain immutability. Medical document confidentiality is handled at the application and off-chain storage level, while the blockchain is used for integrity verification, consent management, and auditability. Encryption keys are not stored on-chain, which prevents the blockchain layer from becoming a repository of sensitive or directly exploitable medical information. Security mechanisms are integrated directly into application flows, including hash-based tamper detection and on-chain verification of access rights. The prototype is evaluated through realistic operational scenarios, analyzing security properties, performance, and transaction cost implications. Results show that, relative to a DB-only baseline, the hybrid approach provides structurally stronger support for integrity verification, traceability, and accountability without exposing sensitive medical data on-chain. The study also highlights practical limitations related to latency and costs in public blockchain environments, supporting a selective on-chain design focused on high-value operations.**

Keywords*—*Electronic Health Records (EHR); blockchain, ethereum; smart contracts; access control; consent management; auditable access

I. Introduction

Currently, the digitalization of the medical field through EHR is no longer just a problem of computerization, but one of data architecture and governance. In this context, the benefits of EHRs are real, but they are conditioned by the ability of the systems to securely manage sensitive data, to control access in a transparent way and to provide traceability of the actions performed on the data.

Blockchain represents a technology that proposes a distributed ledger in which transactions are recorded in a verifiable form, and the history is difficult to alter without the consensus of the network [1]. The concept was initially popularized by Bitcoin as a mechanism for maintaining a ledger resistant to unauthorized changes in the absence of a central authority [2]. Later, platforms such as Ethereum extended this paradigm by introducing smart contracts – programs that run on the blockchain and can implement business rules and access mechanisms in an automated and transparent way [3].

This study aims to analyze how blockchain technology can support the security and management of electronic health records (EHR), in a practical solution in which blockchain is used as a governance, evidence, and audit layer. From an implementation perspective, each document is registered through the smart contract by storing its hash and associated reference, while a corresponding event is emitted to create an auditable proof of registration. Access control is handled through explicit grant and revoke operations, which update the on-chain permission state and emit audit events for traceability. When a document is requested, the application verifies the current access permission through the smart contract before allowing the off-chain encrypted content to be retrieved and displayed. The approach was supported by the implementation of a functional prototype with a hybrid architecture (web application developed in Next.js, off-chain storage via Prisma/DB and on-chain component on Ethereum in a local test environment), used as a case study to practically validate the basic hypothesis: medical data (content) must be stored off-chain, and blockchain must be used selectively for integrity, consent and traceability, without exposing the medical content. Choosing Ethereum allows for the analysis of practical advantages and limitations of a blockchain solution, based on different criteria such as transaction costs, performance, and considerations regarding on-chain vs. off-chain data storage. The main contributions of this work are as follows:

1) We propose a hybrid Ethereum-based architecture for electronic health records (EHR) that separates sensitive medical data storage from the trust and verification layer. In this architecture, medical data is stored off-chain, while Ethereum is used for governance, integrity anchoring, consent management, and auditability.

This work was funded by the University of Oradea, Romania

2) We introduce a consent-driven access control model implemented through smart contracts, enabling patient-centric authorization workflows across multiple stakeholders. The prototype supports explicit access granting and revocation operations, and each relevant action is recorded through blockchain events to provide an auditable history.

3) We provide a practical design for integrity anchoring of medical records using cryptographic hashes stored on-chain. Each medical document can be linked to an on-chain hash, allowing later verification that the off-chain content has not been modified, while avoiding the exposure of sensitive medical data on the blockchain.

4) Unlike prior work that focuses mainly on permissioned blockchain environments such as Hyperledger Fabric, our approach evaluates the feasibility of using Ethereum as a deployment environment. The work includes a cost-oriented analysis based on gas consumption and varying gas price assumptions, highlighting the economic constraints of public blockchain deployment.

5) We present a prototype implementation and evaluate its performance characteristics in a local Ethereum-based testing environment. The evaluation highlights the difference between local testing conditions, where transaction execution is fast and cost-free, and public blockchain deployment, where latency, gas costs, and scalability constraints must be considered.

6) We discuss the architectural implications of key management and off-chain encryption for secure real-world deployment. In the implemented prototype, encryption and sensitive data handling are kept outside the blockchain layer, while the blockchain stores only hashes, references, permissions, and audit-relevant events. This separation reduces the exposure of sensitive medical information and supports a more GDPR-aligned system design.

The study is structured as follows: Section II presents the related work and the implementation gap that motivates the proposed architecture; Section III presents the architecture and security requirements of the Blockchain (Ethereum)-based EHR prototype; Section IV describes the security issues and their integration into the application functional flows; furthermore, a comprehensive analysis and discussions regarding advantages and limitations based on several criteria such as security, performance, scalability and costs, were presented in Section V. Finally, some conclusions were drawn.

II. Related Work and Identified Research Gap

Existing research on blockchain-based EHR systems can be grouped into three main directions: conceptual frameworks and literature reviews, prototype-oriented implementations, and studies focused on the trade-offs between on-chain and off-chain storage. The first group, represented by surveys and framework papers [4]-[9], confirms the relevance of blockchain for healthcare interoperability, privacy and auditability, but most contributions remain at an architectural or conceptual level and do not describe an executable application flow that combines encrypted off-chain storage, smart-contract consent and hash-based integrity verification in a single prototype.

A second group contains practical implementations, including solutions based on permissioned environments such as Hyperledger Fabric [10] or more general blockchain-based record-management prototypes [11]. These works are valuable because they show that blockchain can be integrated with healthcare workflows, but they often assume a controlled consortium setting, focus mainly on record registration, or do not explicitly analyze the cost implications of Ethereum-style public deployment. As a result, the specific operational tension between public-chain transparency, transaction fees, medical confidentiality, and GDPR-oriented data minimization remains insufficiently addressed.

The third direction discusses the general storage trade-off between keeping data on-chain and storing only proofs or references off-chain [12]. This literature motivates the hybrid design adopted in this study. However, the identified limitation is that the storage trade-off is frequently discussed independently from concrete EHR consent flows. The proposed prototype addresses this gap by combining three mechanisms in the same implemented case study: encrypted off-chain medical documents, on-chain consent state through grant/revoke operations, and on-chain integrity anchoring through cryptographic hashes and audit events.

Therefore, the purpose of this work is not to replace existing EHR platforms or to claim production-level scalability from a local prototype, but to validate a concrete hybrid security pattern and to analyze its practical implications. The novelty of the study lies in the integrated implementation and evaluation of consent, integrity anchoring, and auditable access in an Ethereum-compatible setting, while explicitly keeping sensitive medical content outside the blockchain layer.

III. Case Study – Blockchain (Ethereum) - Based EHR Prototype

A. Prototype Architecture

To analyze and evaluate a blockchain-based EHR solution, a functional Ethereum-based EHR prototype was developed in a local environment to analyze, in a reproducible way, how security advantages such as integrity, auditability, and consent-based access control can be obtained, as well as investigating the problems that such implementations raise from a practical implementation perspective.

The approach, based on a realistic design, tries to reduce exposure and avoid storing clinical content in raw form on a replicated ledger; for this reason, the prototype uses a hybrid architecture: off-chain for content (document content + operational metadata - local database via Prisma ORM [13]) and on-chain for governance and integrity (access rules, integrity proofs - hashes and audit trails - events). This pattern, frequently recommended in healthcare literature, balances confidentiality with verifiability [14].

The prototype is built around three functional modules, visible in the application interface as separate sections and reflected in the overall architecture presented in Fig. 1, which illustrates the explicit separation between the off-chain and on-chain areas as a main security requirement of the project:

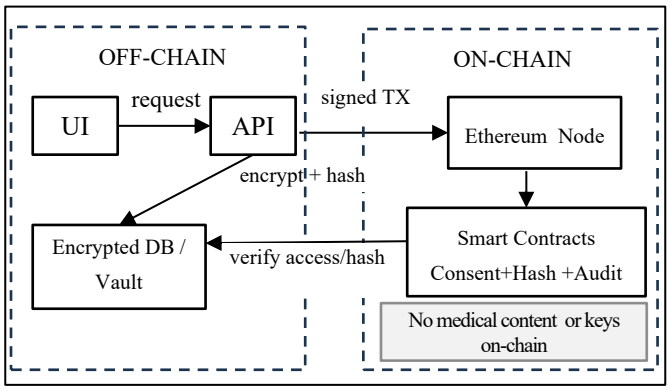


Fig. 1. Overall architecture with off-chain/on-chain storage.

- Document Vault (off-chain): upload, encrypted storage, metadata, indexing, document listing.
- Consent Manager (on-chain + application): grant/revoke access for doctors, with verifiable status on-chain.
- Audit Explorer (on-chain + application aggregation): visualization of events issued by smart contracts (registered document, granted/revoked access), as an "append-only" chronological log.

From the implementation perspective, the prototype is composed of several components. The web client (Next.js UI) is represented by an interface logically separated by roles (patient/doctor). UI offers: uploading documents (patient), listing own documents (patient), granting/revoking access to a doctor (patient), accessing allowed documents (doctor), viewing the audit (both roles, with different levels). Backend (Next.js server: API routes/server actions) performs security logic such as input validations, encryption/decryption, hash calculation, DB interaction (Prisma), calls to smart contracts (local Ethereum). Local blockchain (Ethereum in our test environment) implements contracts for permissions (grant/revoke), recording hashes, issuing audit events. Off-chain storage (DB + Prisma) keeps encrypted content, keeps metadata necessary for searching/listing, and keeps the link between an off-chain document and its on-chain proof (e.g. docId + hash + txHash / blockNumber).

B. Functional and Security Requirements

The functional requirements of the prototype include the following: register users with distinct roles (e.g. patient, doctor); upload medical documents (e.g. results, reports, medical letters) and associate them with the patient; grant access by the patient to a doctor (grant), with the possibility of revocation; consult documents by the doctor only if there is active permission; maintain an audit trail for critical actions (upload, grant, revoke, consult).

Security requirements (what the system must protect) are represented by:

- Confidentiality - medical content must not be exposed publicly; access must be conditioned and controlled.
- Integrity - any unauthorized modification of documents stored off-chain must be detected.
- Auditability / non-repudiation - relevant actions must be traceable (who granted access, when it was revoked, etc.).
- Data minimization - only strictly necessary data is stored on the blockchain (hashes, technical identifiers, events).
- Separation of responsibilities - neither the blockchain nor the database are sufficient on their own; security is achieved end-to-end (application + DB + contracts + keys).

1. User
id (UUID)
role (PATIENT / DOCTOR)
displayName
ethAddress (0x… associated with the wallet account)
createdAt
2. Document
id (UUID – internal docId)
ownerUserId (pacient)
type (ex: analyses, imaging, letter, recipe)
createdAt (moment upload)
ciphertext (blob / encrypted file) or internal pointer if storing in filesystem
cipherAlgo (ex: AES-256-GCM)
hashAlgo (ex: SHA-256)
contentHash (consistently calculated hash)
sizeBytes
anchorTxHash (tx on blockchain)
anchorBlockNo
status (ACTIVE / TAMPERED / REVOKED_BY_OWNER)
3. PermissionCache (optional, for UI performance)
patientUserId
doctorUserId
statusLocal (informational mirror)
lastCheckedAt

Fig. 2. Entity conceptual model.

The off-chain component is represented by a local database accessed through Prisma ORM. The reason for this choice is twofold: from a performance perspective, it ensures efficiency for searches, listings, and file/metadata storage, and from a security perspective it ensures increased confidentiality, with the medical content remaining in a controllable area, where classic policies can be applied (backup, encryption, access control, retention). In the prototype, the off-chain modelling aims to clearly separate the following elements:

- Application identity data: users, roles, mapping to an Ethereum address (for signing transactions).
- Medical documents: content (stored encrypted or in protected form), metadata (type, date, issuer).
- Access bindings: a local representation of permissions (optional), used for UX and caching, but the final decision is made based on the on-chain state.
- Application audit: useful internal logs (e.g., errors, requests), separate from the on-chain audit (which covers critical actions).

The on-chain model uses Ethereum/Ganache for governance, permissions, and audit; the component is represented by a set of smart contracts implemented in Solidity [15] and run on a local Ethereum network (Ganache [16]). Their

role is to be a verifiable registry for: access permissions (grant/revoke – formal decision, signed by the patient through a transaction); integrity proofs (the document hash, anchored in the registry); audit events – critical actions are reflected through events, so that the audit can be rebuilt.

The design of the contracts follows a strict principle: no medical content or explicit clinical information is stored on-chain. The contract only contains elements such as technical identifiers, such as document IDs, hashes, Ethereum addresses, logical timestamps, and events. It can be seen that the key to the approach is minimization: in the DB, only what is necessary for operation and security is kept, and in the blockchain, only the evidence/documents and rules are kept.

IV. Security Issues and Implementation into Functional Flows

Both on-chain and off-chain layers require specific approaches from the security point of view for implementation:

A. Security in the Off-Chain Data Model (Database)

In the prototype, the DB has the role of enforcing a series of security invariants: the medical content is stored encrypted; there is no "direct read" of the content without on-chain verification; the document is associated with a hash and with the reference to the on-chain proof; permissions are doubly conditional: on-chain state + integrity validation.

A conceptual model of tables/entities (compatible with our Next.js + Prisma implementation) is presented in Fig. 2.

The model clearly separates the identity and roles of users, the protected medical content, and the metadata necessary for operation, respectively, the verifiable link to the on-chain proofs. Thus, the medical content remains off-chain and encrypted (confidentiality), the integrity is verifiable through anchored hashes (integrity), and critical actions are linked to on-chain transactions and events (auditability and traceability).

It is noted that the User entity contains only the minimum information necessary for operation (id, role, display name) and a blockchain layer connection field (ethAddress), used for mapping between the application account and the address that signs transactions. Choosing this field allows auditability and non-repudiation for critical operations (grant/revoke/register), while keeping the detailed clinical identity off-chain (data minimization principle).

The Document entity is designed so that the medical content remains off-chain, protected: the ciphertext field (or an internal pointer to the encrypted file) allows the data to be stored in encrypted form, and the cipherAlgo and hashAlgo fields explicitly document the algorithms used (e.g., AES-256-GCM, SHA-256), increasing reproducibility and verifiability. The contentHash field represents the consistently calculated fingerprint of the content and is the key to the integrity mechanism: upon access, the recalculated hash from the DB is compared with the anchored value. The anchorTxHash and anchorBlockNo fields directly link the off-chain document to the on-chain transaction in which the proof was recorded, allowing independent verification of the moment and existence of the proof.

The status field (ACTIVE/TAMPERED/REVOKED etc.) is introduced to reflect the security state observed by the application (e.g. hash mismatch → TAMPERED) and to prevent displaying a suspicious document.

The PermissionCache entity is optional and has a strictly UI performance role: it can locally store the result of the last on-chain check (per patient–doctor–document), together with lastCheckedAt, to reduce repeated calls to the contract in fast navigation scenarios. Importantly, this entity is not the source of truth for permissions; the final access decision remains on-chain (hasAccess), and the cache is only an information mechanism that can be revalidated/invalidated based on events (AccessGranted/AccessRevoked) or upon expiration.

B. Security and On-Chain Layer: Smart Contracts and Executable Rules (consent + Audit)

In the prototype, the blockchain is used strictly as a governance and evidence layer. For clarity and a small attack surface, it is recommended to use a two-contract design.

Contract 1 – ConsentRegistry (used for permissions) has the following responsibilities: stores the state of permissions: patient → doctor → allowed/denied; optional: includes expiration (timestamp) and a minimal "scope" (document types). Typical functions are the following: grantAccess (address doctor, uint64 expiresAt) – called by the patient; revokeAccess(address doctor) – called by the patient; hasAccess(address patient, address doctor) – view (used by the server). Typical audit events: AccessGranted(patient, doctor, expiresAt); AccessRevoked(patient, doctor).

Contract 2 – DocumentAnchor (used for hashes), has the following responsibilities: records the proof of integrity for documents: (patient, docId) → hash; emits a DocumentAnchored event on each record. Typical functions: anchorDocument(bytes32 docId, bytes32 contentHash) – called by the patient (or by an authorized "uploader" on their behalf); getHash(address patient, bytes32 docId) – view. Typical events: DocumentAnchored(patient, docId, contentHash).

A justified design decision in this regard is that the events are intentionally minimal: they do not include diagnoses, results, document names, or other sensitive metadata, containing only technical identifiers and actions (grant/revoke/anchor).

C. Integrating Security Issues into Implementation Flows

One of the challenges of the implementation was to integrate security elements as smoothly as possible into the functional flows of the application.

Flow A – Document upload, off-chain protection, and on-chain anchoring of proof of integrity is shown in Fig. 3.

- Step 1 (UI – patient): patient uploads a .pdf file, e.g., analyses
- Step 2 (server): backend validates file type and size; generates a per-document symmetric key; encrypts the content (e.g., AES-256-GCM); calculates a hash (e.g., SHA-256) over encrypted content (or over original content – it is important to be consistent); saves in DB: ciphertext, contentHash, docId, metadata.

- Step 3 (on-chain): application sends transaction anchorDocument(docId, contentHash) to contract; contract records hash and issues DocumentAnchored.
- Step 4 (DB completion): DB is updated with anchorTxHash and anchorBlockNo.

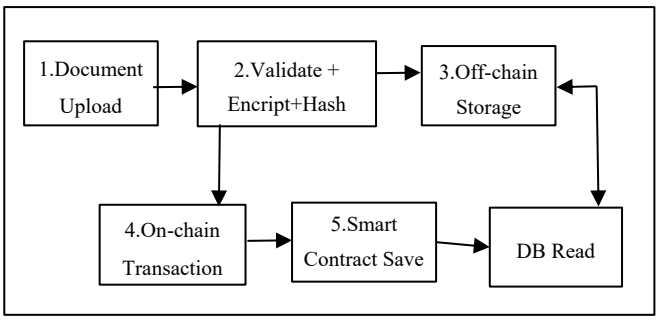


Fig. 3. Document upload, off-chain protection, and on-chain anchoring of proof of integrity.

Flow B – Granting access and document consultation sequence (permission + integrity check) is shown in Fig. 4.

- Step 1 (UI – patient): patient selects a doctor (e.g. 0xB2...) and sets duration (e.g. 7 days).
- Step 2 (on-chain): transaction grantAccess(doctor, expiresAt) is sent; contract updates permission state and issues AccessGranted.
- Step 3 (UI/Audit Explorer): event appears in log: AccessGranted – patient X → doctor Y.

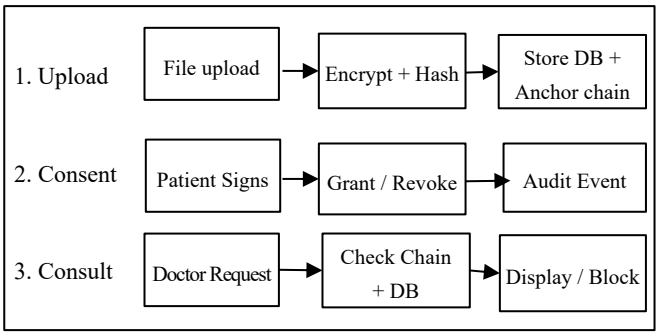


Fig. 4. Granting access and document consultation sequence (permission + integrity check).

Thus, in the prototype, a grant is a cryptographically signed transaction (patient wallet) and produces an on-chain event, which makes consent easier to prove and harder to challenge later (who, to whom, for what and when).

Flow C – doctor's document consultation (access control + integrity), involves the following steps:

- Step 1 (UI – doctor): the doctor requests the document docId = 7b2f.
- Step 2 (server): the application runs: hasAccess(patient, doctor) - view on-chain → if false, stops; retrieves ciphertext from DB; recalculates the hash on the content (same method as for upload); extracts the on-chain hash from getHash(patient, docId); compares hashes: if they do not match → status = TAMPERED, access blocked + alert; if everything is valid → decrypts and displays the content.

As an observation related to security, it should be mentioned here that the blockchain does not "keep" the document, but acts as a log: any subsequent change becomes detectable. At the time of upload, the document's cryptographic fingerprint (hash, e.g. SHA-256) is calculated, and this value is anchored on-chain (via registerDocument and/or the DocumentRegistered event). Subsequently, at each consultation, the application recalculates the hash on the document retrieved from off-chain storage and compares it with the anchored hash; if the file has been modified, replaced, or corrupted in the database, the recalculated hash will be different, and the discrepancy becomes an objective indicator of alteration, causing access to be blocked/decryption denied or the document to be marked as invalid.

Flow D – Access Revocation involves the following steps:

- Step 1 (UI – patient): the patient revokes the doctor's access.
- Step 2 (on-chain): a transaction revokeAccess(doctor) is made; an AccessRevoked event is generated.
- Step 3 (practical effect): any further consultations are blocked at Step 2 of scenario C (permission check).

An inherent limitation in this context is that revocation does not "erase" the fact that the doctor has already seen the document. It only stops future access; therefore, auditing and organizational policies remain necessary.

Flow E – Tamper detection: in this scenario, if an attacker (or operational error) modifies the ciphertext in the DB, then:

- upon access, the recalculated hash no longer matches the hash anchored on-chain;
- the application marks the document as TAMPERED;
- blocks decryption and display.

This is exactly where the prototype offers a structural advantage over a traditional DB-only implementation: there is an external, immutable proof of the "original state" (the hash anchored on-chain) that cannot be rewritten by accessing the off-chain infrastructure. In a traditional system, an attacker or a privileged administrator who compromises the DB can, at the same time, modify the document, update the metadata (including any hash stored internally), and adjust the application logs to reduce the chances of detection.

In the prototype implementation approach, even if the file in the DB is replaced or edited, the hash recalculated upon consultation will no longer coincide with the value recorded on-chain, and the discrepancy becomes an objective technical indicator of the alteration (tamper evidence). To "hide" the alteration, the on-chain evidence (the anchored hash) would also have to be modified, which is not possible retroactively without a new transaction visible in the history and, implicitly, without leaving audit trails (timestamp/block number). Therefore, the blockchain transforms integrity from a trust-based property into a verification-based property, where the document is accepted

only if it passes cryptographic validation reported to the on-chain anchor.

D. Privacy and GDPR Compliance

The proposed system operates in the context of electronic health records (EHR), which are classified as special category data under Article 9 of the General Data Protection Regulation (GDPR) [17]. As such, privacy protection and data minimization are fundamental requirements of the architecture.

A key challenge arises from the conflict between blockchain immutability and the "right to erasure" under Article 17 GDPR [17]. The proposed architecture mitigates this issue by ensuring that no raw personal data or medical records are stored directly on-chain. Instead, the blockchain stores only cryptographic hashes, document references, access permissions, and audit-relevant events, while the actual medical content is kept encrypted in off-chain storage. As a result, sensitive off-chain data can be deleted, modified, or made inaccessible when required, without needing to alter immutable blockchain records.

The system also supports a consent-driven authorization model. Patient consent is represented through smart contract-based access control operations, such as granting and revoking access to a medical document. These operations create a verifiable audit trail through blockchain transactions and events, allowing the system to demonstrate when access was granted, revoked, or checked.

Data minimization principles under Article 5 GDPR are addressed by limiting the information stored on-chain to technical verification data [17]. No direct identifiers, raw medical records, diagnoses, treatment details, or other sensitive health information are stored on the blockchain. The on-chain hash is used only as an integrity proof, allowing the application to verify whether the corresponding off-chain document has been altered.

Regarding cross-border data transfer under Article 44 GDPR, the use of public Ethereum infrastructure may introduce additional compliance risks, since blockchain nodes can be distributed across multiple jurisdictions [17]. For a real-world deployment, this risk could be mitigated by using controlled access gateways, EU-based node providers, private or consortium Ethereum-compatible networks, or infrastructure operated by trusted healthcare institutions.

Finally, encryption and key management remain critical for confidentiality. In the prototype, confidentiality is handled outside the blockchain layer: medical documents are stored off-chain in encrypted form, while the blockchain is used for integrity verification, access control, and auditability. Encryption keys are not stored on-chain, reducing the risk that immutable blockchain records could expose sensitive or directly exploitable medical information

E. Key Management and Data Protection

The security of the proposed system depends not only on the smart contract logic, but also on the protection of the off-chain medical data and the cryptographic material used to secure it. Since the blockchain is not used to store medical content directly, the confidentiality of the system relies on the encryption and access-control mechanisms implemented at the application and off-chain storage layers.

In the implemented prototype, medical document confidentiality is handled outside the blockchain layer. The blockchain stores only hashes, references, access permissions, and audit events, while the sensitive document content is stored off-chain. This separation ensures that the blockchain can provide integrity verification and auditability without becoming a repository of confidential medical information.

Encryption keys are not stored on-chain. This is an important design decision because any data written to a blockchain should be considered persistent and difficult or impossible to remove. Keeping keys outside the blockchain prevents unauthorized users from deriving access to medical content directly from blockchain data.

For a production-ready implementation, a more advanced key management model would be required. One possible approach is to encrypt each medical document using a document-specific symmetric key and then protect that key using the public keys of authorized users. Under such a model, only users with the corresponding private keys would be able to decrypt the medical data. However, this introduces additional requirements related to secure private key storage, recovery mechanisms, access revocation, and user usability.

Revocation of access is also an important practical consideration. In the prototype, access revocation is represented at the smart contract level by updating the on-chain permission state. This prevents a previously authorized user from passing future access checks through the application. However, in a real-world system, stronger revocation guarantees may require re-encryption of the off-chain data and redistribution of new keys to the remaining authorized users.

Therefore, the architecture assumes that blockchain-based authorization must be combined with secure off-chain key lifecycle management. Future improvements may include integration with hardware-backed keystores, hardware security modules, decentralized identity frameworks, or institutional key management services suitable for healthcare environments.

For a production deployment, the key lifecycle would need to be specified in more detail. Key rotation could be implemented through envelope encryption, where each document key is wrapped by user or institutional public keys and can be re-wrapped without rewriting the encrypted document itself. Multi-device access would require controlled enrollment of additional user keys, while emergency recovery or break-glass access would require explicit institutional policy, strong authentication, and additional audit events. Institutional delegation and insider threats should be handled through role separation, least privilege, hardware-backed key storage, and periodic access reviews. In practice, these controls would normally be implemented through a healthcare-grade KMS/HSM or institutional identity infrastructure rather than directly through the blockchain.

While the proposed architecture improves integrity, auditability, and transparency, several security considerations must be addressed before real-world deployment.

One potential risk is the use of application-level caching mechanisms for access permissions. If permission data is cached, there is a possibility of stale-state inconsistencies, for example when access has been revoked on-chain but the application still relies on an outdated cached permission state. This risk can be reduced through short cache lifetimes, explicit cache invalidation after grant or revoke operations, and periodic revalidation against the smart contract state.

Another important consideration is the reliance on off-chain storage. Since medical documents are not stored directly on-chain, the system depends on the availability and security of the external storage layer. This introduces trust assumptions regarding storage infrastructure, backup policies, and operational security. However, the impact of this risk is reduced by encrypting the off-chain content and verifying document integrity against the hash anchored on-chain.

The correctness of the smart contract logic is also essential. Functions responsible for document registration, access granting, revocation, and permission checking must be carefully designed to prevent unauthorized access or inconsistent authorization states. Since smart contract transactions are immutable once deployed, contract testing and validation are critical before deployment to a production blockchain environment.

Finally, the overall security of the system depends heavily on cryptographic key protection. If a user's private key or application-level encryption material is compromised, unauthorized access to medical documents may become possible. Therefore, secure key storage, recovery mechanisms, user awareness, and institutional security policies are necessary components of any real-world deployment.

V. Analysis and Discussions

A. Security Analysis and Evaluation

The security analysis based on the prototype case study is structured on the following criteria: confidentiality, integrity, availability, complemented by access control and auditability.

1) Threat model and security invariants: To make the security evaluation more explicit, the prototype can be interpreted under the following threat model: an attacker may attempt to read or modify off-chain stored documents, an unauthorized doctor may attempt to access a patient document without consent, a stale application cache may preserve an outdated permission state, or a privileged infrastructure actor may attempt to alter local database records or logs. The model does not claim to solve denial-of-service attacks, theft of a user's private key, or all smart-contract implementation bugs; these remain deployment risks that require additional operational controls.

The main security invariants enforced by the prototype are: I1 - raw medical content and encryption keys are never written on-chain; I2 - a document is displayed only if hasAccess(patient, doctor) returns an active permission state; I3 - the hash recalculated from the off-chain content must match the hash anchored on-chain before decryption/display; I4 - PermissionCache is never the source of truth and cannot override the smart contract state; I5 - grant, revoke and anchor operations generate auditable events that can be reconstructed independently from the application database.

Against these invariants, the defense mapping is concrete: database tampering is detected by the mismatch between the recalculated hash and the on-chain anchor; unauthorized access is blocked by the hasAccess check; stale-cache risk is mitigated by short cache lifetime, event-based invalidation and revalidation against the contract; modification of centralized logs does not remove the blockchain event history. Smart contract vulnerabilities are reduced by a small contract surface and testable state transitions, but formal verification and third-party audit would be required before production deployment.

Confidentiality is achieved by: off-chain storage of content; encryption at the application level (before DB); decryption conditional on on-chain verification + integrity. The specific blockchain risk (in public networks) is leakage through metadata and address correlations. In the prototype (local), the risk is low, but for a real system, mitigation measures should be mentioned: minimization of events, pseudonymization, avoiding including the on-chain document type, aggregation, and rate-limiting at the UI/API level. However, blockchain solution privacy remains dependent on encryption and key management (off-chain). Blockchain does not guarantee privacy by itself; it only reinforces it through a design that minimizes what is public.

Integrity is strengthened by on-chain anchored hash and verification at each access. This mechanism is simple, cheap to implement, and very strong as "proof": any difference between the stored content and the anchored proof becomes immediately visible at the time of access, without relying exclusively on centrally managed logs. From an integrity perspective, there is a clear advantage of the blockchain solution: integrity becomes verifiable through an external proof (on-chain hash), not just through trust in the administration.

Regarding access control and consent (non-repudiation and accountability), as shown in the prototype, consent is not a setting in the DB, but a verifiable state on-chain. This produces two effects: access becomes independently verifiable (at least at the system level) and actions become traceable and hard to deny (transaction signature). Grant/revoke actions are cryptographically signed transactions (address/key), and the change of permission state is recorded on-chain. This results in an important practical property: stronger non-repudiation for critical actions (who granted/revoked access and when). In addition, the prototype applies "double checking": even if the DB contains the document, actual access is conditional on on-chain permission verification + hash integrity verification. Thus, from an access control perspective, the blockchain solution reduces the risk of "invisible" abuse at the consent level and makes access decisions easier to prove later.

Audit is built on on-chain events (grant/revoke/anchor). The advantage over centralized logs is the append-only nature and transaction signing. Basically, even if the DB is compromised, the main history of critical actions remains reconstructable from the chain: on-chain events (DocumentAnchored, AccessGranted, AccessRevoked, etc.) are append-only and can be used to reconstruct a verifiable timeline of critical actions. Consequently, from the audit perspective, there is a real

auditability gain, especially for critical actions, in the hybrid blockchain model, as the on-chain log is harder to tamper with without a trace.

Because the content is still off-chain, the availability of documents still depends on the DB and the application. Blockchain does not replace storage and does not solve availability on its own. Thus, the hybrid blockchain approach does not eliminate the off-chain dependency: if the DB goes down, the documents are not accessible, even if the evidence is on the chain. Therefore, in a real architecture, availability remains largely a classic problem and relies on classical approaches (backup, replication, monitoring).

2) Qualitative comparison with a DB-only EHR baseline: The baseline considered for the security comparison is a conventional centralized EHR implementation in which the document, the stored hash, permission flags, and audit logs are all maintained inside the same administrative database or infrastructure domain. In such a baseline, a privileged compromise can modify the medical file and also rewrite the associated hash, permission record, or local audit trail. In the proposed hybrid architecture, the attacker would additionally have to change the previously anchored on-chain hash or remove historical grant/revoke events, which is not possible retroactively without creating a new visible transaction. Therefore, the improvement claimed in this study is a structural and verifiability-oriented improvement, not a quantitative throughput comparison against an independently implemented EHR system.

B. Performance and Scalability Analysis

The performance of a hybrid EHR (off-chain + on-chain) cannot be evaluated as a single "response time", because the flows have two distinct components: 1) off-chain operations, which dominate the user experience (upload, read, list), and 2) on-chain operations, which dominate the governance decision (consent, anchor, audit).

1) Measurement methodology: The measurements were performed in the local environment of the prototype, using the following approach: measured end-to-end time for each operation (from the action in the UI to the response received by the user); separation of times by stages, where relevant: 1) server processing (validation + encryption + hash), 2) DB operation (Prisma), 3) blockchain interaction (submit tx + confirmation), 4) access checks (hasAccess + getHash + compare).

For on-chain operations, two different values were measured: t_submit (how long it takes to send the transaction, until you receive txHash); t_confirm (how long it takes until the transaction is mined/confirmed and the event is available). This separation is important because the UX can be designed so that the user does not "wait in a blocking way" for confirmation in some cases (e.g., anchoring can run asynchronously), but for other cases confirmation is critical (e.g., grant/revoke must be confirmed to be considered active).

2) Example measurements on the prototype case-study: Table I summarizes average values measured on a set of common files and on typical scenarios (upload, grant, consultation).

TABLE I. AVERAGE TIMES PER PROTOTYPE (MS) – END-TO-END MEASUREMENT (LOCAL ENVIRONMENT)

Operation	Measurement results		
	Dominant Components	***Average Time (ms)***	***Observations***
Upload document 1 MB (off-chain)	encryption + DB write	~180–320	does not include on-chain confirmation
Hash calculation (SHA-256)1 MB	CPU (server)	~8–15	subdominant to IO
AES-256-GCM 1 MB encryption	CPU (server)	~25–60	depends on implementation
Save DB (Prisma) 1 MB	IO DB	~60–180	depends on local DB
Submit transaction anchor (txHash)	RPC + signing	~120–250	user sees "sent"
Confirm transaction (Ganache)	local mining	~400–900	stable in controlled environment
GrantAccess (submit)	RPC + signing	~120–250	similar to anchor
GrantAccess (confirm)	local mining	~400–900	becomes active after confirmation
Consult document (verify + display) 1 MB	hasAccess + DB read + compare + decrypt	~220–520	includes on-chain checks (view)

The values are indicative for the local environment and serve to highlight where costs arise and which components dominate. From measurements, we can notice that for off-chain operations, the cost is dominated by I/O (DB) and encryption. Hash is fast and does not represent a significant bottleneck. Also, for on-chain operations, confirmation is the most "visible" component if it is put directly into the clinical flow (the user waits). In the local prototype, confirmation is stable; in real public networks, this latency is variable and can increase significantly. Document consultation remains comparable to a traditional system in terms of access time, with a moderate overhead due to checks (hasAccess + getHash + compare), but with a security gain (integrity + verifiable access control).

3) Interpretation of local measurements and deployment limits: The measurements in Table I must be interpreted as prototype-level measurements obtained in a local Ganache environment. They validate the correctness of the implemented flows and identify the dominant components inside the prototype, but they do not prove public Ethereum feasibility. Ganache eliminates variable consensus latency, fee volatility,

and network congestion; therefore, the public deployment discussion in this study is limited to gas-based cost estimation and architectural implications. A real deployment would require additional testing on public test networks, layer-2 networks, or a controlled consortium network.

The prototype was not stress-tested with thousands of patients, concurrent doctors, or very large repositories. In a production architecture, high-frequency actions should remain off-chain, while on-chain operations should be limited to high-value state transitions such as document anchoring and consent updates. Operational resilience would also require transaction queues, retry and idempotency logic for pending blockchain operations, redundant blockchain RPC providers or nodes, database replication and backups, monitoring, and a fail-closed access policy when the system cannot verify current permission or integrity state.

Storage overhead introduced by the hybrid model is limited compared with the medical files themselves. For each document, the additional off-chain fields include an internal document identifier, content hash, hash/encryption algorithm identifiers, initialization vector and authentication tag for encryption, transaction hash, block number and status flags. These artifacts are typically below a few kilobytes per document, while the medical document can be several megabytes. The on-chain footprint is dominated by the stored hash/reference and event data, while detailed audit records and application logs should remain off-chain or be periodically anchored to avoid high costs.

The current evaluation uses fixed representative parameters, such as a 1 MB document and a moderate number of annual grant/revoke/access operations. A complete sensitivity analysis should vary document size, document frequency, transaction frequency, cache lifetime, encryption overhead, number of users, and concurrent requests. This limitation is acknowledged because the present study focuses on validating the security pattern and estimating cost order of magnitude, not on exhaustive capacity planning.

C. Costs Analysis

1) Premises for comparison (annual scenario): The cost estimations were made on a moderate scenario, realistic enough to estimate their order of magnitude. The assumptions on which the analysis was based are presented in Table II.

TABLE II. PREMISES FOR COST ASSESSMENT (ANNUAL SCENARIO, PER PATIENT)

Parameter	Value (scenario)	Observations
New documents / patient / year	10	e.g.: tests, imaging, medical letter
Average document size	1 MB	order of magnitude for common files
Stored volume / patient / year	10 MB	10 doc × 1 MB
Document requests / year	20	doctor requests for access
Access grants / year	4	access grants (consent)
Access revokes / year	4	revoke access

In public Ethereum, each on-chain operation has a transaction cost (fee), calculated by:

$$\text{fee (ETH)} = \text{gasUsed} \times \text{gasPrice} \quad (1)$$

where, gasPrice is in gwei (1 gwei = 10^-9 ETH).

In the local prototype (Ganache), this monetary cost is practically 0, but in a real scenario (public Ethereum) the cost becomes relevant. For example, we use a further estimation valid at present of price_eth = 2,212.07 EUR/ETH. Plausible estimates for minimalist contracts (mapping + events) were the following: document anchor (hash + event): 95,000 gas; grant access (grant + event): 60,000 gas; revoke access (revoke + event): 45,000 gas; log access (optional audit event): 35,000 gas. GasPrice scenarios (so as not to be "cosmetic numbers") could be chosen from the following: 10 gwei (low), 30 gwei (medium), 60 gwei (high). Consequently, the direct formula that we used will be:

$$\text{Cost (EUR)} = \text{gasUsed} \times \text{gasPrice(gwei)} \times 10^{-9} \times \text{price_eth (EUR/ETH)} \quad (2)$$

2) Costs comparison results: It can be seen from Table III that, in public Ethereum, there are significant cost per access, making a model where "everything is logged on-chain" unsustainable. This leads to the correct conclusion: on public networks, blockchain is suitable for rare high-value operations (anchors, critical consent), not for turning every access into a transaction. Therefore, for real adoption, some architectures preserve integrity/auditability, but drastically reduce the cost. One of the options is represented by off-chain audit + periodic on-chain anchoring or networks (predictable operational cost).

TABLE III. HYBRID EHR COST ESTIMATION ON PUBLIC ETHEREUM (10, 30, 60 GWEI VARIANTS)

Parameter	Costs Public Ethereum /year		
	10 gwei	***30 gwei***	***60 gwei***
10 document anchors/year/pacient	21 EUR	63 EUR	126 EUR
20 document requests / year (audit)/pacient	15,40 EUR	46,40 EUR	92,80 EUR
4 Access grants / year	5,32 EUR	15,92 EUR	31,84 EUR
4 Access revokes / year	4,00 EUR	11,96 EUR	23,92 EUR
TOTAL/pacient/year	**45,72 EUR**	**137,28 EUR**	**274,56 EUR**
TOTAL for 10000 pacients	**457.200 EUR**	**1.372.800 EUR**	**2.745.600 EUR**

VI. CONCLUSION

This study demonstrates that a hybrid blockchain architecture provides a robust framework for Electronic Health Records (EHR), where integrity stands out as the most direct and demonstrable benefit. The developed prototype confirms that blockchain serves as an essential tamper-evident mechanism, making any modification, replacement, or corruption of documents within the database objectively detectable. Through this approach, consent and access control transcend being internally modifiable states; they become a transparent set of actions with verifiable traces, documenting precisely who granted access, to whom, for which document, and when.

Regarding regulatory compliance and data protection, the architecture was designed to support GDPR-oriented requirements by enforcing a strict separation between sensitive medical data and immutable blockchain records. The implemented prototype does not store raw personal information or medical content on-chain. Instead, the blockchain layer stores only cryptographic hashes, document references, access permissions, and audit-relevant events, while the actual medical documents remain encrypted and stored off-chain. This design supports the principle of data minimization and reduces the risk of exposing sensitive health information through immutable blockchain records.

Privacy is therefore not treated as an inherent blockchain feature, but as an end-to-end architectural responsibility. The blockchain is used for integrity verification, consent management, and auditability, while confidentiality is handled at the application and off-chain storage layers. During document retrieval, the application verifies the current on-chain access permission and the integrity of the requested document before allowing the encrypted off-chain content to be accessed and displayed. Encryption keys are not stored on-chain, which prevents the blockchain layer from becoming a source of sensitive or directly exploitable medical information.

The feasibility of this model depends on a selective on-chain design. To mitigate the high costs and variable fees associated with public Ethereum transactions, only the critical minimum—cryptographic evidence, consent updates, and essential events—is recorded on-chain. Detailed auditing and intensive data operations remain off-chain to maintain a controllable cost profile and competitive user experience (UX) performance.

While the current evaluation was conducted in a local Ganache environment, the prototype provides a functional proof of concept for the proposed architecture, including document registration, hash-based integrity anchoring, access granting, access revocation, and permission verification. The local environment was suitable for validating the correctness of the smart contract logic and the interaction between the application layer and the blockchain layer, but it does not fully capture the operational characteristics of a public Ethereum network.

Consequently, the results should be interpreted as a functional proof of concept and as an order-of-magnitude cost and security analysis. They should not be read as evidence that the same performance would be obtained on public Ethereum or under large healthcare workloads without further load testing, operational hardening, and production-grade key management.

Further research will also explore the adoption of permissioned or consortium networks for increased identity control and the integration with medical standards such as FHIR/HL7. Ultimately, this hybrid approach proves that blockchain can effectively serve as a specialized consent and audit layer, seamlessly integrating with existing healthcare infrastructures while ensuring security, accountability, and legal compliance.

REFERENCES

[1] S. Khezr, M. Moniruzzaman, A. Yassine and R. Benlamri, “Blockchain Technology in Healthcare: A Comprehensive Review and Directions for Future Research”, Appl. Sci. 2019, 9(9), 1736; https://doi.org/10.3390/app9091736

[2] Nakamoto, S. “Bitcoin: A Peer-to-Peer Electronic Cash System”. 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[3] Buterin, V. “Ethereum White Paper: A Next-Generation Smart Contract and Decentralized Application Platform”. 2014. [Online]. Available: https://ethereum.org/en/whitepaper/

[4] Y. Han , Y. Zhang and S. H. Vermund, “Blockchain Technology for Electronic Health Records”, Int J Environ Res Public Health. 2022 Nov 24;19(23):15577. doi: 10.3390/ijerph192315577

[5] N. Ettaloui, S. Arezki and T. Gadi, “Blockchain-Based Electronic Health Record: Systematic Literature Review”, 2024, https://doi.org/10.1155/hbe2/4734288

[6] Blockchain for Healthcare, IEEE Xplore Digital Library, November 20, 2025 [Online]. Available: https://innovate.ieee.org/innovation-spotlight/blockchain-for-healthcare/

[7] A. Althaf Ali, M.A. Gunavathie, V. Srinivasan, M. Aruna, R. Chennappan, M. Matheena Securing electronic health records using blockchain-enabled federated learning for IoT-based smart healthcare, Clinical eHealth, Volume 8, December 2025, Pages 125-133

[8] N. Ul. A. Tahir, U. Rashid, H. J. Hadi, N. Ahmad, Y. Cao, M. A. Alshara and Y. Javed, “Blockchain-Based Healthcare Records Management Framework: Enhancing Security, Privacy, and Interoperability”, Technologies 2024, 12(9), 168; https://doi.org/10.3390/technologies12090168

[9] K. Pampattiwar and P. Chavan, “A secure and scalable blockchain-based model for electronic health record management”. Sci Rep 15, 11612 (2025). https://doi.org/10.1038/s41598-025-94339-w

[10] O. A. Oki, A. O. Agbeyangi and A. Mgidi, “Blockchain in Healthcare: Implementing Hyperledger Fabric for Electronic Health Records at Frere Provincial Hospital” [Online]. Available: https://arxiv.org/html/2407.15876v1

[11] S. Gajarlewar, A. Patle, A. Bhosale, B. Khamkar and A. Sase (2026). “HealthVault: Trustworthy Medical Records on the Blockchain”. In: I. Bhiradi, J. Machado, (eds) Intelligent Control, Robotics, and Industrial Automation. RCAAI 2024. Lecture Notes in Electrical Engineering, vol 1492. Springer, Singapore.

[12] H. Eren, Ö. Karaduman and M. T. Gençoğlu, “Security Challenges and Performance Trade-Offs in On-Chain and Off-Chain Blockchain Storage: A Comprehensive Review”, Appl. Sci. 2025, 15(6), 3225; https://doi.org/10.3390/app150632

[13] Prisma ORM. Next-generation Node.js and TypeScript ORM [Online]. Available: https://www.prisma.io/orm

[14] T. T. Kuo, H.E. Kim, L. Ohno-Machado, “Blockchain distributed ledger technologies for biomedical and health care applications”. Journal of the American Medical Informatics Association, Nov 1;24(6):1211-1220. doi: 10.1093/jamia/ocx068, 2017

[15] The Solidity Authors, “Introduction to Smart Contracts”. [Online]. Available: https://docs.soliditylang.org/en/latest/introduction-to-smart-contracts.html

[16] Truffle Suite. Ganache – Documentation. [Online]. Available: https://archive.trufflesuite.com/docs/ganache/

[17] General Data Protection Regulation GDPR. [Online]. Available: https://gdpr-info.eu/

Using Artificial Intelligence Techniques for Error Detection and Repair in Software Development

Abdulaziz Aladwani, Sultan Alsamaani, Turki Alrumaykhani, Mohamed Tahar Ben Othman
Department of Computer Science-College of Computer, Qassim University, Qassim, Saudi Arabia

***Abstract*—Software bugs remain one of the most costly challenge in software engineering, consuming significant development time and resources. Recent advances in Artificial Intelligence (AI), particularly deep learning and large language models (LLMs), have shown remarkable potential in automating the detection and repair of software errors. This study presents a comprehensive survey and comparative analysis of AI-based techniques for error detection and automated program repair (APR). We categorize existing approaches into traditional search-based methods, learning-based neural machine translation models, and emerging LLM-based repair systems. We evaluate these techniques across standard benchmarks, including Defects4J and SWE-bench, comparing their effectiveness in terms of bugs fixed, patch correctness, and scalability. Our analysis reveals that while LLM-based approaches significantly outperform traditional methods in repair capability, challenges remain in patch correctness validation, computational cost, and generalization to real-world codebases. Our results show that LLM-based tools, such as ChatRepair, can correctly fix 114 out of 395 benchmark bugs at just $0.42 per fix, fixing 2.6× as many bugs as the best traditional method (a 165% increase). We discuss open challenges and propose future research directions toward more reliable AI-assisted software development.**

Keywords*—Automated program repair; bug detection; deep learning; large language models; software engineering; neural machine translation; defect prediction; code analysis

I. Introduction

Software defects are an inherent part of the software development lifecycle, costing the global economy an estimated $1.56 trillion annually, according to the Consortium for Information and Software Quality (CISQ). Debugging and fixing software errors account for a substantial portion of development effort, with studies estimating that developers spend 35–50% of their time on debugging activities. The consequences of undetected bugs range from minor inconveniences to catastrophic failures: the 2014 Heartbleed vulnerability in OpenSSL affected an estimated 17% of all secure web servers, while the 2017 Equifax breach—caused by a single unpatched vulnerability—exposed the personal data of 147 million individuals. As software systems grow in complexity, with modern applications spanning millions of lines of code across distributed architectures, the need for automated approaches to detect and repair errors becomes increasingly critical.

The traditional software debugging workflow follows a manual process: a developer identifies a failing test or user report, localizes the fault through debugging tools and code inspection, designs a fix, implements the patch, and validates it against the test suite. This process is time-consuming, error-prone, and does not scale with the pace of modern software development where thousands of commits are made daily in large organizations.

Artificial Intelligence has emerged as a transformative force in software engineering, offering new paradigms for automating tasks traditionally performed by human developers. The application of AI to error detection and repair spans multiple generations of techniques: from early search-based approaches like GenProg [1] that use genetic programming to evolve patches, to modern large language models that can understand and generate code with remarkable fluency [2]. The key insight driving this progress is that software bugs follow recurring patterns—a study of over 700 real-world Java bugs found that approximately 25% involve conditional statement errors, 20% involve method call changes, and 15% involve variable replacements [3].

Such single-statement errors—for instance, a bisection method in the Apache Commons Math library (Defects4J, Math-70) that returns its initial endpoint instead of the computed root, illustrated concretely in Fig. 3 of Section III—are precisely the kind of recurring pattern that AI techniques can learn to detect and correct.

The evolution of AI-based program repair can be broadly characterized in three waves:

- Search-based and constraint-based repair (2009–2017): Techniques such as GenProg [1], Prophet [4], and Nopol [5] that search for patches using test suites as correctness oracles.
- Learning-based repair (2018–2021): Neural machine translation models like SequenceR [6], CoCoNuT [7], and CURE [8] that learn bug-fix patterns from large code corpora.
- LLM-based repair (2022–present): Pre-trained language models, including CodeBERT [9], Codex [2], and ChatGPT-based systems [10] that leverage massive pre-training for zero-shot or few-shot repair.

Similarly, AI-driven bug detection has progressed from traditional machine learning classifiers for defect prediction [11] to sophisticated deep learning models that analyze program semantics through graph neural networks [12] and pre-trained code representations [13].

This study makes the following contributions:

- A comprehensive taxonomy of AI techniques for software error detection and repair, covering 32 key works from 2012 to 2025.

- A comparative analysis of traditional, learning-based, and LLM-based approaches across standard benchmarks.
- An identification of open challenges and future research directions in AI-assisted software debugging.

Fig. 1 illustrates the taxonomy of AI techniques for error detection and repair covered in this survey.

The remainder of this study is organized as follows: Section II reviews related work. Section III provides theoretical background on the key AI techniques. Section IV describes our survey methodology. Section V presents our comparative case studies. Section VI discusses results and findings. Section VII discusses threats to validity. Section VIII concludes with future work directions.

II. Related Work

A. Traditional Automated Program Repair

The field of automated program repair was pioneered by GenProg [1], which introduced genetic programming to evolve patches by mutating program statements guided by test suite fitness. While groundbreaking, GenProg's patches often lacked semantic correctness. Prophet [4] addressed this by learning probabilistic models from successful human patches to prioritize correct repairs. Angelix [14] introduced symbolic analysis for scalable multiline patch synthesis, while Nopol [5] specialized in repairing conditional statement bugs through constraint-based synthesis.

Template-based approaches further improved repair quality. SimFix [15] leveraged existing patches and similar code at AST granularity, fixing 34 bugs on the Defects4J benchmark. TBar [3] revisited template-based repair with systematically refined fix templates, achieving 43 bug fixes on Defects4J.

B. Learning-Based Program Repair

The application of neural machine translation (NMT) to program repair marked a paradigm shift. SequenceR [6] was among the first end-to-end sequence-to-sequence models for repair, incorporating a copy mechanism to handle out-of-vocabulary tokens. DLFix [16] employed a two-tier tree-based RNN to learn contextual code transformations, fixing 2.5× more bugs than previous baselines.

CoCoNuT [7] combined context-aware NMT models using ensemble methods with a multi-stage attention mechanism. CURE [8] pre-trained NMT models on large code corpora and integrated static checking to ensure generated patches contain valid identifiers.

Industrial deployments validated the practical impact of learning-based repair. Getafix [17], deployed at Facebook, used hierarchical clustering of fix patterns from static analysis violations, demonstrating that learning-based repair can operate at production scale in industrial environments.

C. LLM-Based Error Detection and Repair

The emergence of large language models has substantially advanced the state-of-the-art. Xia and Zhang [18] demonstrated that LLMs can repair 3.3× more bugs than prior approaches through zero-shot cloze-style repair (AlphaRepair). ChatRepair [10] leveraged conversational LLMs with iterative feedback, fixing 162 out of 337 Defects4J bugs at $0.42 per fix. Chen et al. [19] taught LLMs to self-debug by explaining and correcting their own generated code. Most recently, RepairAgent [20] introduced autonomous agent-based repair with planning and iterative patch refinement.

For vulnerability-specific repair, VulRepair [21] fine-tuned T5 models to fix 745 out of 1,706 real-world CVE vulnerabilities. RewardRepair [22] incorporated reinforcement learning with execution-based feedback to guide neural repair models.

D. AI-Based Bug and Vulnerability Detection

DeepBugs [23] pioneered learning-based bug detection by training neural networks on identifier names to detect swapped arguments and wrong operators. For vulnerability detection, Devign [12] applied graph neural networks to learn comprehensive program semantics, while VulDeePecker [24] used bidirectional LSTMs on code gadgets extracted through program slicing.

Pre-trained code models significantly advanced detection capabilities. CodeBERT [9] provided the first bimodal pre-trained representation for code and natural language. GraphCodeBERT [13] incorporated data flow information into pre-training, improving code understanding tasks. LineVul [25] achieved 160–379% higher F1-measure using Transformer-based architectures for line-level vulnerability prediction.

E. Software Defect Prediction

Machine learning has been extensively applied to predict defective software modules. Yang et al. [11] introduced deep belief networks for just-in-time defect prediction from code commits. Transfer learning approaches [26] enabled cross-project defect prediction by aligning feature distributions across different software projects.

F. Existing Surveys

Several surveys have systematically reviewed this field. Gazzola et al. [27] provided a comprehensive taxonomy of APR techniques. Hou et al. [28] reviewed 395 papers on LLMs for software engineering. Our work differs by providing an integrated comparative analysis spanning detection and repair, with a focus on the recent LLM-based paradigm shift.

III. Theoretical Background

Fig. 2 shows the chronological evolution of key techniques across the three waves of AI-based program repair, with bug/vulnerability *detection* milestones marked by a dagger (†) to distinguish them from the repair techniques.

A. Formal Foundations of Program Repair

Automated program repair can be formally defined as follows. Given a buggy program P_b and a test suite $T = \{t_1, t_2, \ldots, t_n\}$ where at least one test $t_f \in T$ fails on P_b, the goal is to produce a patched program P_p such that all tests in T pass:

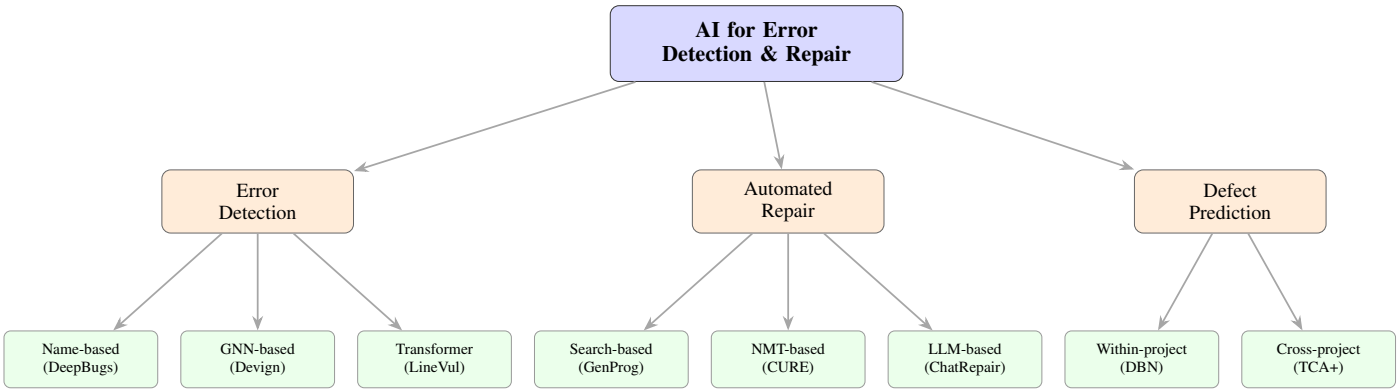


Fig. 1. Taxonomy of AI techniques for software error detection and repair.

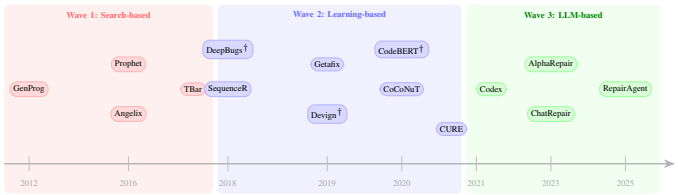


Fig. 2. Evolution of AI-based program repair and detection (†) techniques (2012–2025). Repair techniques are organized into three waves. Detection milestones (DeepBugs, Devign, CodeBERT), marked with †, are shown on the same axis for chronological context only and do not constitute repair "waves".

$$\text{APR}(P_b, T) = P_p \text{ such that } \forall t_i \in T : t_i(P_p) = \text{pass} \quad (1)$$

This formulation connects to the theory of computation through the concept of program equivalence and the undecidability of determining semantic correctness from a finite test suite (the *overfitting problem*).

To make this formulation concrete, Fig. 3 shows a representative single-statement repair from Defects4J (Math-70): a bisection method that incorrectly returns its initial endpoint instead of invoking the root-finding computation. This is precisely the class of recurring single-statement error that an AI-based repair tool can detect and correct automatically.

```
// Buggy version:
return initial;     // <-- always returns
                    //     the initial guess
// Fixed version (by AI repair):
return solve(f, min, max);  // <-- computes
                            //     the root
```

Fig. 3. Example of bug fix from Defects4J Math-70: a single-line change that an LLM-based tool can detect and repair automatically.

B. General APR Workflow

Algorithm 1 presents the general workflow shared by most automated program repair systems, regardless of the underlying technique used to generate candidate patches.

Algorithm 1 Conversational Automated Program Repair

Require: Buggy program P_b, test suite T, patch generator G, max iterations N
Ensure: Repaired program P_p or failure

```
1:  loc ← FAULTLOCALIZATION(P_b, T)
2:  for each suspicious location l ∈ loc do
3:      ctx ← ∅ {conversational feedback context}
4:      for iter ← 1 to N do
5:          candidates ← G(P_b, l, ctx) {G conditioned on feedback}
6:          for each patch p ∈ candidates do
7:              P_p ← APPLY(P_b, p)
8:              result ← RUNTESTS(P_p, T)
9:              if result = all pass then
10:                 return P_p
11:             else
12:                 ctx ← ctx ∪ DIAGNOSTICS(result) {errors, failing tests}
13:             end if
14:         end for
15:     end for
16: end for
17: return failure
```

The key differentiator between approaches is the patch generator G: search-based methods use genetic programming, NMT-based methods use encoder-decoder neural networks, and LLM-based methods use pre-trained language models. Crucially, the outer iteration loop (lines 4–5) captures the multi-turn, conversational behavior of modern LLM-based and agentic tools such as ChatRepair [10] and RepairAgent [20]: after each failed validation attempt, diagnostic feedback—compiler error messages and failing-test traces—is accumulated in the context ctx and fed back into the generator, allowing it to refine subsequent patches. Traditional search- and template-based methods are recovered as the special case $N = 1$, where the generator receives no execution feedback (a passive generate-and-validate loop).

C. Search-Based Approaches

Search-based APR techniques explore the space of possible patches using metaheuristic algorithms. GenProg [1] models

repair as a search over program variants using genetic programming, where mutations (insert, delete, replace) are applied to AST nodes and fitness is measured by test suite passing rate.

D. Neural Machine Translation for Code

Learning-based APR formulates repair as a translation task from buggy code to fixed code:

$$P(\text{fix} \mid \text{bug}) = \prod_{i=1}^{m} P(y_i \mid y_1, \ldots, y_{i-1}, \mathbf{x}) \tag{2}$$

where, $\mathbf{x}$ is the buggy code sequence and y_i are tokens of the fixed code, generated autoregressively by an encoder-decoder architecture.

E. Transformer Architecture and Pre-training

The Transformer [29] architecture, based on self-attention mechanisms, has become the foundation for modern code models. The core self-attention mechanism computes weighted relationships between all token pairs in the input sequence, enabling the model to capture long-range dependencies in code—such as matching a variable declaration with its usage hundreds of lines later. Pre-trained models like CodeBERT [9] learn bidirectional code representations through masked language modeling (MLM) on large code corpora. This enables zero-shot and few-shot capabilities for downstream tasks including bug detection and repair.

F. Graph Neural Networks for Code Analysis

Graph neural networks model code as graphs (e.g., control flow graphs, data flow graphs, ASTs) to capture structural properties. Devign [12] processes composite code property graphs through gated graph neural network layers, where each node iteratively aggregates information from its neighbors using a GRU-based update function. This graph-based formulation allows the model to reason about program semantics beyond the linear token sequence, capturing control dependencies, data flow relationships, and calling conventions that are critical for accurate vulnerability detection.

G. Reinforcement Learning for Program Repair

Reinforcement learning (RL) offers a complementary paradigm to supervised learning for program repair. In the RL formulation, the repair agent interacts with an environment (the program and its test suite) by generating candidate patches (actions) and receiving feedback based on test execution results (rewards). RewardRepair [22] uses a policy-gradient approach where the neural repair model is trained to maximize expected rewards from test execution, directly optimizing for test-passing patches rather than merely imitating human-written fixes. This approach has been extended by training critic models that predict patch correctness before execution, enabling more efficient exploration and early filtering of invalid candidates.

H. Transfer Learning and Domain Adaptation

Transfer learning has proven essential for addressing the data scarcity problem in software error detection. Within-project defect data is often limited, making it difficult to train accurate models. Transfer defect learning [26] addresses this by using Transfer Component Analysis (TCA+) to learn a transformation matrix that aligns the feature distributions of source and target projects in a shared feature space, minimizing the domain discrepancy while preserving the discriminative structure of the data. Pre-trained code models like CodeBERT [9] implicitly perform transfer learning by learning general-purpose code representations that can be fine-tuned for specific tasks with minimal labeled data.

IV. Methodology

This section describes the systematic methodology adopted for conducting this comparative survey, following established guidelines for systematic literature reviews in software engineering.

A. Research Questions

This study aims to address the following research questions:

- RQ1: How do traditional, learning-based, and LLM-based APR techniques compare in terms of bugs fixed and patch correctness?
- RQ2: What are the strengths and limitations of AI-based bug detection approaches across different vulnerability types?
- RQ3: How do repository-level, agent-based systems perform on real-world issue resolution (SWE-bench), and how does this compare with isolated-bug benchmarks?
- RQ4: What are the key challenges and future directions for AI-assisted error detection and repair?

B. Search Strategy and Paper Selection

We conducted a systematic search across four major digital libraries: IEEE Xplore, ACM Digital Library, Springer Link, and arXiv. The search was performed using the following query structure:

(“automated program repair” OR “bug detection” OR “vulnerability detection” OR “defect prediction”) AND (“deep learning” OR “neural” OR “machine learning” OR “language model” OR “transformer”)

The initial search returned over 2,500 candidate papers. We applied the following inclusion and exclusion criteria to filter relevant works:

Inclusion criteria:

- Published in top-tier SE/AI venues (ICSE, FSE, ASE, ISSTA, NeurIPS, ICLR, ICML, TSE, TOSEM, OOPSLA)
- For papers published in 2012–2022, a minimum of 50 citations was required as an impact threshold. For

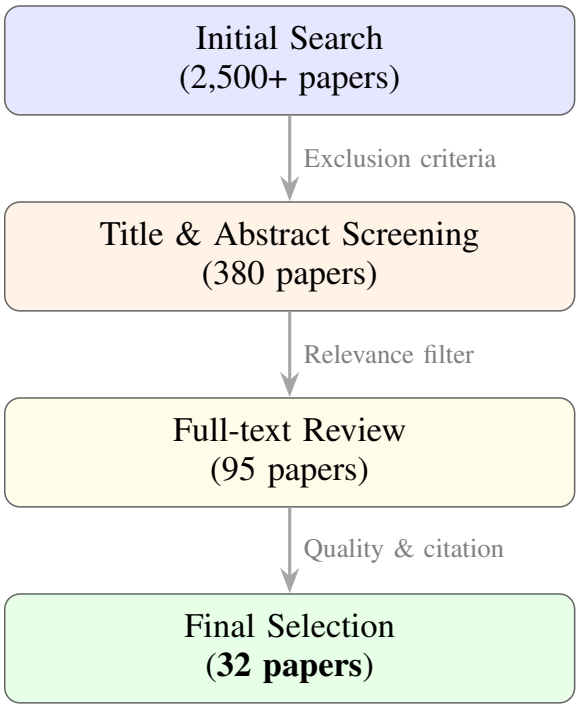


Fig. 4. Systematic paper selection process (PRISMA-inspired flow).

TABLE I. COMPARISON OF DIMENSIONS AND EVALUATION METRICS

Dimension	Metrics
Effectiveness	Bugs fixed, precision, recall, F1 score
Scalability	Training time, inference latency, memory usage
Generalizability	Cross-project, cross-language performance
Correctness	Ratio of correct to plausible patches
Practicality	Industrial deployment, cost per fix, ease of integration

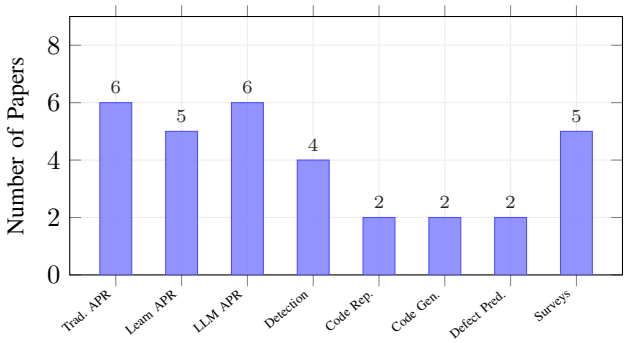


Fig. 5. Distribution of the 32 selected papers across research categories.

recent papers (2023–2025), the citation requirement was waived—since recent work cannot plausibly have accumulated comparable citations—and inclusion was instead based on venue quality, methodological novelty, and relevance, in order to capture emerging trends

- Primary focus on AI/ML/DL techniques for error detection, localization, or repair
- Published between 2012 and 2025

Exclusion criteria:

- Papers focusing solely on test generation without bug detection/repair
- Short papers, tool demonstrations, or posters under 6 pages
- Papers not evaluated on established benchmarks
- Duplicate or extended versions of already-included works

After applying these criteria, screening titles and abstracts, and performing full-text review, we selected 32 papers for detailed analysis. Fig. 4 illustrates the paper selection process.

C. Comparison Framework

We compare techniques across five key dimensions, each capturing a different aspect of practical utility. Table I summarizes these dimensions and their associated metrics.

D. Data Extraction and Analysis

For each selected paper, we extracted: 1) the AI technique category, 2) the training data source and size, 3) the evaluation benchmarks used, 4) quantitative results on standard metrics, and 5) reported limitations. We organized findings by technique category and synthesized cross-cutting observations to answer our research questions.

We emphasize that this work is a *secondary* study (a comparative literature survey): we did not train, fine-tune, or re-evaluate any model ourselves, and consequently performed no data cleaning, class balancing (e.g., SMOTE), or train/test partitioning. The only "pipeline" in this study is the paper-selection process of Fig. 4. All quantitative results reported in Section V and Section VI are taken verbatim from the respective primary studies under their original experimental protocols; questions of data leakage or resampling order therefore pertain to those primary works rather than to this survey. Where primary studies use incomparable configurations, we flag this explicitly in Section VII.

Fig. 5 shows the distribution of selected papers across the eight categories defined in our taxonomy.

V. CASE STUDIES AND COMPARATIVE ANALYSIS

A. Benchmark: Defects4J

Defects4J [30] is the most widely used benchmark for evaluating APR techniques, containing 835 real bugs from 17 open-source Java projects. Table II compares key approaches on Defects4J v1.2 (395 bugs).

Fig. 6 visualizes the comparison between correct and plausible patches across the three categories.

B. Benchmark: SWE-bench

SWE-bench [31] evaluates LLMs on 2,294 real GitHub issues from 12 popular Python repositories including Django,

TABLE II. APR PERFORMANCE ON DEFECTS4J V1.2 (395 BUGS)

Technique	Category	Correct	Plausible
GenProg	Search	5	27
Nopol	Constraint	5	22
SimFix	Template	34	56
TBar	Template	43	81
SequenceR	NMT	14	21
DLFix	NMT	30	46
CoCoNuT	NMT	28	44
CURE	NMT	44	57
AlphaRepair	LLM	80	110
ChatRepair	LLM	**114**	**162**

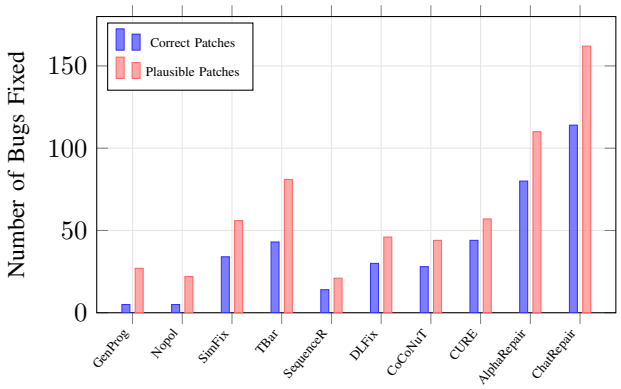


Fig. 6. Comparison of APR techniques on Defects4J v1.2: correct vs. plausible patches across three categories.

Flask, scikit-learn, and sympy. Unlike Defects4J which provides isolated single-function bugs with failing tests, SWE-bench presents realistic software engineering challenges that require: 1) understanding natural language issue descriptions, 2) navigating large codebases to localize the relevant files, 3) implementing multi-file patches that maintain consistency across the codebase, and 4) passing existing regression tests.

Table III summarizes the performance of various approaches on SWE-bench Lite (300 representative instances).

TABLE III. PERFORMANCE ON SWE-BENCH LITE (300 INSTANCES)

Approach	Resolved (%)	Year
RAG + GPT-4 (baseline)	1.96	2024
SWE-Agent + GPT-4	12.47	2024
AutoCodeRover	19.00	2024
Aider + GPT-4o	26.33	2024
RepairAgent	32.00	2025

Fig. 7 visualizes the rapid improvement on SWE-bench Lite.

The dramatic gap between isolated bug benchmarks (ChatRepair: 33.8% on Defects4J) and real-world issue resolution (best: 32% on SWE-bench Lite) underscores that real software maintenance requires significantly more than pattern-based patching—it demands holistic codebase understanding.

C. Cost and Efficiency Analysis

A critical practical consideration is the cost-effectiveness of different approaches. Table IV compares the approximate

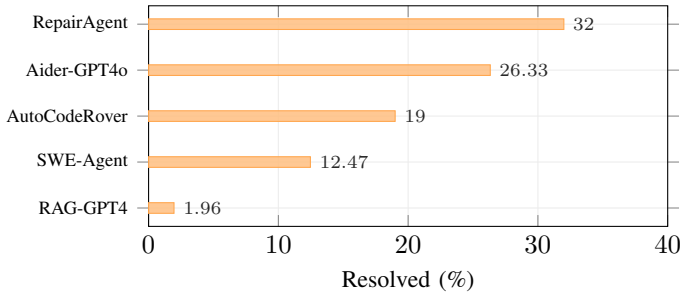


Fig. 7. Resolution rates on SWE-bench Lite (300 real-world GitHub issues).

computational requirements and costs across technique categories.

TABLE IV. COST AND RESOURCE COMPARISON ACROSS APR CATEGORIES.

Category	Training	Inference	Cost/Fix
Search-based	None	Hours	CPU only
Template	None	Minutes	CPU only
NMT-based	GPU days	Seconds	$100+ train
LLM (fine-tuned)	GPU weeks	Seconds	$1000+ train
LLM (API-based)	None	Seconds	$0.42–$5

The API-based LLM approach represents a fundamental shift in the economics of automated repair: by leveraging pre-trained models through APIs, organizations avoid the substantial upfront cost of training while achieving superior repair rates. However, this model introduces dependencies on external service providers and raises concerns about code confidentiality when proprietary code is sent to third-party APIs.

D. Bug Detection Performance

Table V summarizes the performance of key bug detection approaches.

TABLE V. BUG/VULNERABILITY DETECTION PERFORMANCE

Approach	Technique	F1 (%)	Dataset
DeepBugs	Name-based NN	89.0	JS bugs
Devign	GNN	60.1	C/C++ vulns
VulDeePecker	BiLSTM	85.6	NVD vulns
LineVul	Transformer	91.0	Big-Vul
CodeBERT	Pre-trained	62.1	Devign dataset

Fig. 8 provides a visual comparison of bug detection F1 scores.

VI. RESULTS AND DISCUSSION

A. RQ1: Comparative Effectiveness of APR Approaches

Our analysis reveals a clear progression in repair capability across the three generations of APR techniques. As shown in Table II and Fig. 6, LLM-based approaches (ChatRepair: 114 correct fixes) dramatically outperform both traditional methods (TBar: 43) and early learning-based methods (SequenceR: 14).

Key findings include:

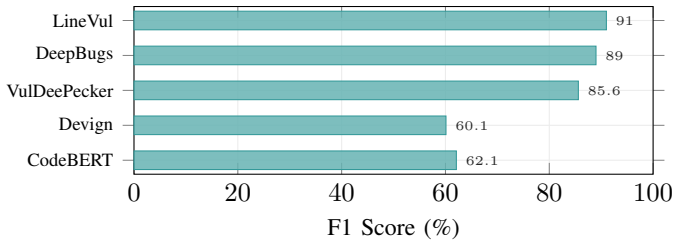


Fig. 8. F1 scores of AI-based bug/vulnerability detection approaches.

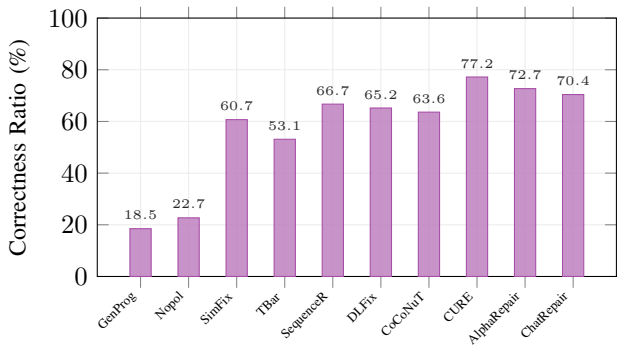


Fig. 9. Patch correctness ratio (correct÷plausible × 100) for each APR technique on Defects4J v1.2.

- LLMs fix 2.6× as many bugs as the best template-based approach on Defects4J (114 vs. 43 correct fixes, a 165% increase), demonstrating the power of large-scale pre-training on diverse code corpora.
- The overfitting problem persists: Even ChatRepair produces 48 plausible but incorrect patches (162 plausible – 114 correct), highlighting that test suite adequacy remains a fundamental bottleneck. This connects to Rice's theorem in the theory of computation—determining semantic correctness from a finite test suite is undecidable in general.
- Cost-effectiveness: ChatRepair's $0.42 per fix represents a paradigm shift in repair economics, making LLM-based repair practical for industrial adoption compared to the estimated $50–$150 cost of manual developer debugging per bug.
- Complementary strengths: Template-based methods (TBar) achieve a higher correct-to-plausible ratio (53%) compared to LLM-based methods (ChatRepair: 70%), suggesting that template methods, while fixing fewer bugs, produce more precise patches within their scope.

Fig. 9 visualizes the patch correctness ratio (correct/plausible) for each technique, revealing how precision varies independently of volume.

An important observation is the diminishing returns within each category. While the jump from Wave 1 (GenProg: 5 fixes) to Wave 2 (CURE: 44 fixes) is an 8.8× increase in correct fixes, the jump from Wave 2 to Wave 3 (ChatRepair: 114 fixes) is only 2.6×, so the rate of improvement is slowing. This suggests that benchmark saturation may be approaching, and future progress may require fundamentally different evaluation methodologies beyond Defects4J.

B. RQ2: Strengths and Limitations of Detection Approaches

Bug detection approaches show complementary strengths, as visualized in Fig. 8. Our analysis identifies a clear trade-off between specialization and generalization:

- Name-based approaches (DeepBugs, F1: 89%) excel at catching surface-level mistakes such as swapped function arguments and incorrect operators. Their key advantage is speed—they operate on identifier embeddings without expensive program analysis. However, they fundamentally cannot detect semantic bugs that do not manifest in naming patterns.
- Sequence-based approaches (VulDeePecker, F1: 85.6%) use bidirectional LSTMs on code slices to capture sequential dependencies. They handle common vulnerability patterns (buffer overflows, use-after-free) well but struggle with bugs that require understanding global program state or inter-procedural data flow.
- GNN-based approaches (Devign, F1: 60.1%) capture deeper program semantics through graph representations of code. While theoretically more expressive, their lower F1 scores reflect challenges with imbalanced datasets (vulnerable functions are rare in practice) and the difficulty of constructing accurate program dependency graphs at scale.
- Transformer-based approaches (LineVul, F1: 91%) achieve the highest detection rates by leveraging pre-trained representations. They provide line-level localization capabilities that are valuable for developer workflows. Their main limitation is the reliance on large pre-training datasets and the potential for distribution shift when applied to programming languages or frameworks not well-represented in the training data.

Fig. 10 plots the maximum number of correct fixes achieved by the best technique in each year, illustrating the accelerating progress driven by each technological wave.

A notable trend is the convergence toward pre-trained models as a foundation for detection tasks. CodeBERT's relatively lower F1 score (62.1%) on the Devign dataset demonstrates that general pre-training alone is insufficient—task-specific fine-tuning with appropriate training objectives (as in LineVul) remains essential for competitive performance.

C. RQ3: Repository-Level Agentic Repair on SWE-bench

As Table III and Fig. 7 shows, agency—not merely model scale—drives repository-level performance. A retrieval-augmented prompting baseline (RAG + GPT-4) resolves only 1.96% of SWE-bench Lite instances, whereas wrapping comparable models in agentic scaffolding that can read files, run tests, and iterate lifts resolution dramatically: SWE-Agent (12.47%), AutoCodeRover (19.00%), Aider + GPT-4o (26.33%), and RepairAgent (32.00%). This is a 16.3× jump

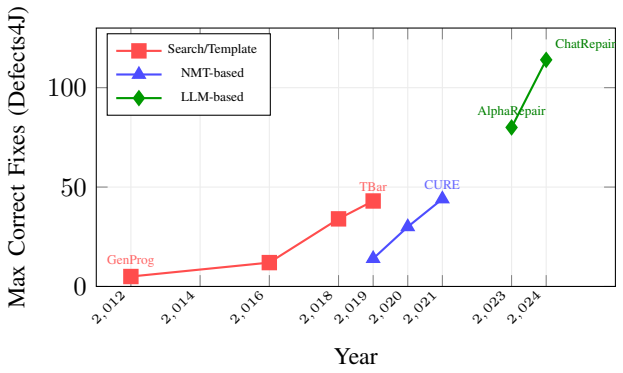


Fig. 10. Progress in maximum correct fixes on Defects4J over time, by technique category.

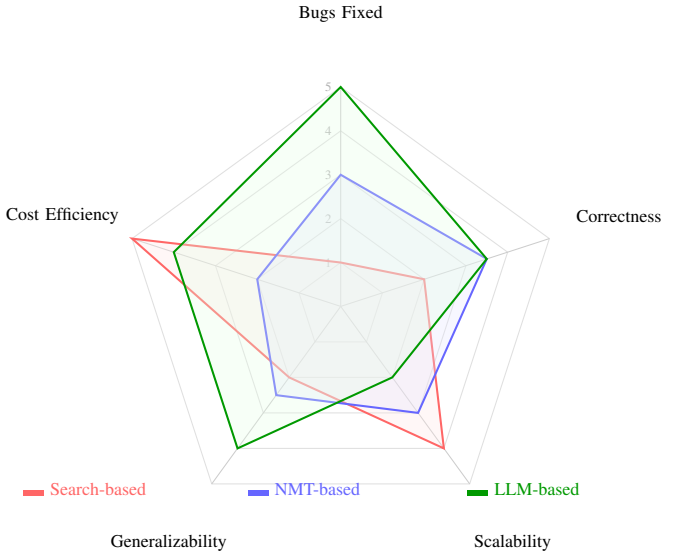


Fig. 11. Multi-dimensional comparison of APR categories across five quality attributes (1=lowest, 5=highest).

from the passive baseline to the best agent, confirming that the ability to interact with the environment—rather than emit a single patch—is the dominant factor at the repository level. This directly mirrors the conversational feedback loop formalized in Algorithm 1.

At the same time, even the best agents lag far behind isolated-bug performance. ChatRepair resolves 33.8% (114/337) of Defects4J bugs it attempts, yet the best agent on SWE-bench Lite reaches only 32%—and does so on issues that are pre-filtered to be "Lite". The gap underscores that fault localization across an unfamiliar codebase, not patch synthesis, is now the binding constraint. We, therefore, treat SWE-bench resolution rate, rather than Defects4J fix count, as the more faithful proxy for practical maintenance capability; the open problems this exposes are analyzed in RQ4.

D. RQ4: Open Challenges

Fig. 11 provides a multi-dimensional comparison of the three APR categories across five key quality attributes, scored on a normalized 1–5 scale based on our analysis.

Despite significant progress, several fundamental challenges remain that represent important directions for the research community:

- Patch Correctness Validation: Distinguishing correct patches from plausible but incorrect ones remains fundamentally difficult. The ratio of plausible-to-correct patches across all techniques averages approximately 1.6:1, meaning nearly 40% of generated patches that pass all tests are semantically incorrect. Without formal verification or significantly more comprehensive test suites, this gap is difficult to close. Emerging approaches include using multiple diverse test suites, semantic equivalence checking, and leveraging LLMs as patch validators.
- Multi-file and Multi-hunk Repair: Most approaches target single-line or single-function fixes within a single file. SWE-bench results (Table III) show that even the best agent-based systems resolve only 32% of real-world issues, many of which require coordinated changes across multiple files. This represents a fundamental challenge because the search space grows combinatorially with the number of edit locations.
- Computational Cost and Accessibility: Training large code models requires significant GPU resources (e.g., StarCoder [32] was trained on 512 GPUs), creating a barrier for academic research groups. While API-based approaches democratize access, they introduce latency, cost, and privacy concerns. Edge deployment of smaller, specialized models remains an active research area.
- Benchmark Quality and Representativeness: Recent studies have identified significant issues with vulnerability detection benchmarks, including mislabeled samples (up to 30% label noise), high duplication rates, and bias toward specific vulnerability types. These issues inflate reported performance and may not reflect real-world detection capability.
- Hallucination and Semantic Correctness: LLMs can generate syntactically valid and test-passing patches that introduce subtle semantic errors—for instance, changing an error message string that coincidentally makes a string-comparison test pass while masking the underlying bug. These "hallucinated" patches are particularly dangerous because they erode developer trust in automated repair systems.
- Explainability: Current AI repair systems operate as black boxes, generating patches without explaining the reasoning behind the fix. For developers to trust and adopt these tools, future systems must provide explanations of why a bug occurred and how the proposed patch addresses the root cause.

VII. Threats to Validity

A. Selection Bias

Our survey covers 32 papers selected from top-tier venues with citation thresholds. This may exclude recent high-quality work from workshops or lower-tier venues, and may over-represent techniques evaluated on popular benchmarks (Defects4J, SWE-bench), while under-representing approaches targeting other languages or domains.

B. Benchmark Limitations

Our comparative analysis relies on results reported in the original papers, which may use different experimental configurations, hardware, and versions of shared benchmarks. Direct comparisons across papers should be interpreted with caution, as subtle differences in fault localization, patch validation, and timeout settings can significantly affect reported numbers.

C. Recency Bias

The rapid pace of LLM development means that some results may already be superseded by the time of publication. We mitigated this by including papers up to 2025, but the field continues to evolve rapidly.

D. Generalizability

Most surveyed techniques are evaluated on Java (Defects4J) or Python (SWE-bench) codebases. Their effectiveness on other programming languages, paradigms (functional, concurrent), or application domains (embedded, real-time) remains largely unvalidated.

VIII. Conclusion and Future Work

This study presented a comprehensive survey and comparative analysis of AI techniques for software error detection and repair. We systematically reviewed 32 key works spanning from 2012 to 2025, categorizing them across three generations of automated program repair (search-based, learning-based, and LLM-based) and multiple categories of bug detection (name-based, sequence-based, graph-based, and transformer-based). Our evaluation used established benchmarks, including Defects4J and SWE-bench.

Our key findings can be summarized as follows:

- LLM-based APR dominates: ChatRepair achieves 114 correct fixes on Defects4J v1.2, fixing 2.6× as many bugs as the best template-based method (TBar: 43) and over 20× as many as early search-based methods (GenProg: 5).
- Real-world gaps remain: SWE-bench performance (best: 32%) reveals that isolated benchmark success does not translate directly to practical software maintenance capability.
- Detection approaches are complementary: No single detection technique dominates across all vulnerability types. Transformer-based methods (LineVul: 91% F1) lead on curated datasets, but specialized approaches remain valuable for specific bug categories.
- Economics favor LLMs: API-based repair at $0.42 per fix fundamentally changes the cost-benefit analysis of automated repair adoption.
- Correctness validation remains the core challenge: Approximately 40% of test-passing patches are semantically incorrect across all technique categories.

A. Future Directions

Based on our analysis, we identify the following promising research directions:

- Agent-based repair systems: Autonomous agents (e.g., RepairAgent [20]) that combine planning, tool use, and iterative refinement represent the most promising direction for complex, multi-file repairs. These systems can navigate codebases, read documentation, run tests, and refine patches iteratively—mimicking the workflow of experienced developers.
- Hybrid detection-repair pipelines: Current systems typically treat detection and repair as separate tasks. End-to-end systems that first localize bugs using GNN-based detection and then apply targeted LLM-based repair could achieve higher accuracy by leveraging the complementary strengths of both approaches.
- Formal verification integration: Connecting LLM-generated patches with lightweight formal methods (e.g., bounded model checking, symbolic execution) could address the correctness validation challenge. A practical approach might use formal methods to filter obviously incorrect patches before test execution.
- Multilingual and cross-platform repair: Most current techniques focus on Java (Defects4J) and Python (SWE-bench). Extending AI repair to languages with different paradigms (Rust, Haskell, Go) and domains (embedded systems, mobile applications) requires new benchmarks and potentially new architectures.
- Developer-in-the-loop systems: Interactive repair systems that present ranked candidate patches with explanations, allowing developers to guide the repair process. This approach combines AI efficiency with human judgment and builds developer trust through transparency.
- Continuous learning from developer feedback: Repair systems that learn from accepted and rejected patches in production environments could continuously improve their accuracy for organization-specific codebases and coding conventions.

B. Implications for Practice

For software development organizations considering AI-based error detection and repair tools, our analysis suggests a practical adoption strategy: begin with LLM-based API tools for simple, single-file bug fixes where the cost-benefit ratio is most favorable; complement with specialized detection tools (e.g., LineVul for vulnerability scanning) in CI/CD pipelines; and invest in comprehensive test suites as the most impactful enabler for all automated repair approaches. As agent-based systems mature, they will increasingly handle the complex multi-file issues that currently require human intervention.

REFERENCES

[1] C. Le Goues, T. Nguyen, S. Forrest, and W. Weimer, "Genprog: A generic method for automatic software repair," *IEEE Transactions on Software Engineering*, vol. 38, no. 1, pp. 54–72, 2012.

[2] M. Chen, J. Tworek, H. Jun, Q. Yuan *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[3] K. Liu, A. Koyuncu, D. Kim, and T. F. Bissyandé, "Tbar: Revisiting template-based automated program repair," in *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2019, pp. 31–42.

[4] F. Long and M. Rinard, "Automatic patch generation by learning correct code," in *Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*. ACM, 2016, pp. 298–312.

[5] J. Xuan, M. Martinez, F. Demarco, M. Clement, S. Marcote, T. Durieux, D. Le Berre, and M. Monperrus, "Nopol: Automatic repair of conditional statement bugs in java programs," *IEEE Transactions on Software Engineering*, vol. 43, no. 1, pp. 34–55, 2017.

[6] Z. Chen, S. Kommrusch, M. Tufano, L.-N. Pouchet, D. Poshyvanyk, and M. Monperrus, "Sequencer: Sequence-to-sequence learning for end-to-end program repair," *IEEE Transactions on Software Engineering*, vol. 47, no. 9, pp. 1943–1959, 2019.

[7] T. Lutellier, H. V. Pham, L. Pang, Y. Li, M. Wei, and L. Tan, "Coconut: Combining context-aware neural translation models using ensemble for program repair," in *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2020, pp. 101–114.

[8] N. Jiang, T. Lutellier, and L. Tan, "Cure: Code-aware neural machine translation for automatic program repair," in *Proceedings of the 43rd International Conference on Software Engineering (ICSE)*. IEEE/ACM, 2021, pp. 1161–1173.

[9] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou, "Codebert: A pre-trained model for programming and natural languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. ACL, 2020, pp. 1536–1547.

[10] C. S. Xia, Y. Wei, and L. Zhang, "Automated program repair via conversation: Fixing 162 out of 337 bugs for $0.42 each using chatgpt," in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2024, pp. 1–13.

[11] X. Yang, D. Lo, X. Xia, Y. Zhang, and J. Sun, "Deep learning for just-in-time defect prediction," *IEEE Transactions on Software Engineering*, 2015.

[12] Y. Zhou, S. Liu, J. Siow, X. Du, and Y. Liu, "Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[13] D. Guo, S. Ren, S. Lu, Z. Feng, D. Tang, S. Liu, L. Zhou, N. Duan, A. Svyatkovskiy, S. Fu, M. Tufano, S. K. Deng, C. Clement, D. Drain, N. Sundaresan, J. Yin, D. Jiang, and M. Zhou, "Graphcodebert: Pre-training code representations with data flow," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[14] S. Mechtaev, J. Yi, and A. Roychoudhury, "Angelix: Scalable multiline program patch synthesis via symbolic analysis," in *Proceedings of the 38th International Conference on Software Engineering (ICSE)*. ACM, 2016, pp. 694–705.

[15] J. Jiang, Y. Xiong, H. Zhang, Q. Gao, and X. Chen, "Shaping program repair space with existing patches and similar code," in *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2018, pp. 298–309.

[16] Y. Li, S. Wang, and T. N. Nguyen, "Dlfix: Context-based code transformation learning for automated program repair," in *Proceedings of the 42nd International Conference on Software Engineering (ICSE)*. ACM, 2020, pp. 602–614.

[17] J. Bader, A. Scott, M. Pradel, and S. Chandra, "Getafix: Learning to fix bugs automatically," in *Proceedings of the ACM on Programming Languages (OOPSLA)*, vol. 3. ACM, 2019, pp. 1–27.

[18] C. S. Xia and L. Zhang, "Less training, more repairing please: Revisiting automated program repair via zero-shot learning," in *Proceedings of the 45th International Conference on Software Engineering (ICSE)*. IEEE/ACM, 2023, pp. 1–13.

[19] X. Chen, M. Lin, N. Schärli, and D. Zhou, "Teaching large language models to self-debug," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[20] I. Bouzenia, P. Devanbu, and M. Pradel, "Repairagent: An autonomous, llm-based agent for program repair," in *Proceedings of the 47th International Conference on Software Engineering (ICSE)*. IEEE/ACM, 2025.

[21] M. Fu, C. Tantithamthavorn, T. Le, V. Nguyen, and D. Phung, "Vulrepair: A t5-based automated software vulnerability repair," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM, 2022, pp. 935–947.

[22] H. Ye, M. Martinez, and M. Monperrus, "Neural program repair with execution-based backpropagation," in *Proceedings of the 44th International Conference on Software Engineering (ICSE)*. ACM, 2022, pp. 1–12.

[23] M. Pradel and K. Sen, "Deepbugs: A learning approach to name-based bug detection," *Proceedings of the ACM on Programming Languages (OOPSLA)*, vol. 2, pp. 1–25, 2018.

[24] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, and Y. Zhong, "Vuldeepecker: A deep learning-based system for vulnerability detection," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

[25] M. Fu and C. Tantithamthavorn, "Linevul: A transformer-based line-level vulnerability prediction," in *Proceedings of the 19th International Conference on Mining Software Repositories (MSR)*. ACM, 2022, pp. 608–620.

[26] J. Nam and S. Kim, "Transfer defect learning," in *Proceedings of the 35th International Conference on Software Engineering (ICSE)*. IEEE, 2013, pp. 382–391.

[27] L. Gazzola, D. Micucci, and L. Mariani, "Automatic software repair: A survey," *IEEE Transactions on Software Engineering*, vol. 45, no. 1, pp. 34–67, 2019.

[28] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, and H. Wang, "Large language models for software engineering: A systematic literature review," *ACM Transactions on Software Engineering and Methodology*, 2024.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[30] R. Just, D. Jalali, and M. D. Ernst, "Defects4j: A database of existing faults to enable controlled testing studies for java programs," in *Proceedings of the 23rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2014, pp. 437–440.

[31] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan, "Swe-bench: Can language models resolve real-world github issues?" in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[32] R. Li, L. B. Allal, Y. Zi *et al.*, "Starcoder: may the source be with you!" *Transactions on Machine Learning Research*, 2023.

Analyzing User Experience in Mobile Banking Applications Through Text Mining

Bora Lamaj (Myrto)[1], Markela Muça[2], Klodiana Bani[3]
Faculty of Information Technology "Aleksander Moisiu"-Department of Computer Sciences, University of Durres, Albania[1]
Faculty of Natural Sciences-Department of Applied Mathematics, University of Tirana, Albania[2, 3]

***Abstract*—This study attempts to offer a data-driven comprehension of the factors that influence satisfaction and dissatisfaction by examining user reviews, based on a banking mobile application in Albania, specifically the Raiffeisen Bank application. All gathered reviews from Google Play and the App Store have been classified using the multilingual BERT model. It is important to mention that mBERT is particularly suitable given that the reviews are written in both Albanian and English. BERTopic has been used to determine the main topics of classification. After classifying the reviews and creating new dimensions with BERTopic, an ordinal logistic regression model was applied to the generated dataset to assess the predictive power of these dimensions and sentiment polarity on user-assigned satisfaction ratings. This study helps developers and managers to better understand the main factors that influence customer reactions and evaluate the factors that influence the rate of the application and the different bugs they report.**

***Keywords*—Generated reviews; sentiment analysis; topic modeling; BERTopic; mBERT**

I. Introduction

As mobile banking applications increasingly replace traditional branch-based services, user experience has become a critical determinant of customer retention and satisfaction in the banking sector. Beyond serving as transactional platforms, mobile banking applications now present a service delivery mechanism and a point through which banks communicate their value proposition to customers. In this environment, users continuously generate reviews on platforms such as Google Play and Apple Store, which are a source of feedback that indicates user experiences. Unlike controlled survey instruments, these reviews capture real perceptions and recommendations expressed in natural language. The existing literature presents a lack of studies on Albanian sentiment that integrate sentiment analysis, topic modeling, and statistical modeling in a unified framework to analyze the impact of user experience dimensions on satisfaction in our case rating. Most prior research focuses either on sentiment classification or topic extraction in isolation, without systematically linking these dimensions to user satisfaction outcomes. This represents a clear research gap in understanding how different aspects of mobile banking experiences translate into user ratings. This study aims to address this gap in the literature concerning the identification and quantification of experiential dimensions that impact mobile banking application reviews and their relationship to satisfaction levels, which is reflected in star ratings. The research focuses on a second-level commercial bank that operates in Albania. This study focuses on giving answers to the following questions: 1) How does sentiment polarity influence user satisfaction ratings? 2) Which experience dimensions have the strongest impacts on user sentiment and ratings? and 3) Does negative user sentiment exert a stronger effect on ratings than positive sentiment? Based on these questions, the following hypotheses are composed:

H1: Sentiment polarity has a positive effect on user satisfaction ratings.

H2: Technical Issues negatively affect user satisfaction ratings.

H3: Negative reviews have a more significant impact on ratings than positive reviews.

The study contributes to the existing literature in three aspects: First, it applies a state-of-the-art NLP pipeline, combining it with multilingual BERT sentiment analysis and with BER Topic modeling to a real banking dataset. Second, it quantifies the influence of extracted experience dimensions on user ratings through regression analysis [16]. Third, it demonstrates the scalability and interpretability of this approach for continuous reputation monitoring in practical banking management.

The remainder of this study is organized as follows: Section II describes the relevant literature on sentiment analysis, topic modeling, and ordinal approaches to modeling user satisfaction in mobile banking applications. Section III contains details regarding research methodology, including data collection and preprocessing, sentiment classification, BERTopic-based topic modeling, dimensional assignment, and multicollinearity diagnostics. Section IV presents the empirical results, covering sentiment distribution, the identified experience dimensions, and the ordinal logistic regression findings, including a test of H3. Section V discusses these findings in relation to prior literature and outlines the study's limitations. Finally, Section VI presents the conclusions, and Section VII outlines directions for future research.

II. Literature Review

The application of transformer-based models to mobile banking reviews has gained considerable traction in recent years. Regarding topic modeling and definition, [9] provided comparison between BERTopic, Latent Dirichlet Allocation LDA and other probabilistic models [12], [18]. Their results showed that BERTopic produces higher semantic discriminability. These findings were supported also by [2], [5] who validated BERTopic in marketing reviews [8], [10], [20].

The use of topic modeling to understand mobile banking user experience has been explored by several recent studies. The authors [19] applied BERTopic to app store reviews in health data, demonstrating that the framework successfully extracts clinical and meaningful topics from multilingual data. Collectively, these papers suggest that topic modeling pipelines applied to app reviews can produce interpretable dimensions of user experience across different service domains. The existing multilingual BERT (mBERT) model will be suitable for our study by considering that the reviews are in two different languages, English and Albanian. Also, BERTopic was preferred over LDA because BERTopic offers a superior semantic understanding using transformer-based embeddings (BERT), while LDA relies on word co-occurrences. BERTopic is more accurate, and requires less preprocessing and performs better in short text environments such as app reviews [11], [15]. Paper [17] analyzed mobile applications reviews of three countries Bosnia-Herzegovina, Croatia, and Serbia, combining BERT-based with BERTopic modelling and multiple linear regression. It demonstrated that sentiment polarity and technical issues were the strongest predictors of ratings, explaining 58% of the variance. In [9], the authors applied a pipeline, regarding an Indonesian mobile banking application, using sentiment analysis and topic modelling on over 7,000 reviews on Google Play [4]. Their study confirmed that technical reliability and ease of use consistently were the most important dimensions in shaping user perception. Another study by [2] on Canadian banking applications, where reviews were taken from App Store and Google Play, demonstrated login and update errors constituted the primary driver of negative ratings, while usability and feature richness drove positive sentiment.

The present study follows the methodological tradition established by these works but departs from them in one substantive respect. Whereas all reviewed studies treat user ratings as a continuous dependent variable and apply multiple linear regression, this study recognizes that star ratings on a five-point scale constitute an ordinal rather than an interval-level outcome. The distances between consecutive rating categories are not guaranteed to be perceptually or behaviorally equal. A move from one to two stars does not necessarily represent the same magnitude of dissatisfaction as a move from four to five stars. Treating such a scale as continuous violates the assumptions of OLS regression and can produce biased coefficient estimates and misleading inference [2]. To address this, the present study employs Ordinal Logistic Regression with the proportional odds specification, estimating four threshold parameters that separate adjacent rating categories and interpreting predictor effects as changes in the log-odds of falling in a higher rather than lower category. The validity of the proportional odds assumption is confirmed via the Brant test (all $p > .97$), and model fit is evaluated using McFadden R^2 rather than the OLS coefficient of determination. This choice constitutes a methodologically more rigorous treatment of the dependent variable and yields results that are directly interpretable in terms of the ordinal structure of user satisfaction.

III. Framework Methodology

This study integrates natural language processing techniques with regression model to examine how user experience dimensions received from BERTopic provide an authentic feedback reflecting real user experiences and perceptions.

A. Sample Definition and Data Collection

User reviews were collected from the official mobile application of Raiffeisen Bank available on Google Play and App Store. The dataset includes reviews posted from September 2021 to December 2023. A total of 2,764 reviews were retrieved where only 1,000 contained text review, while the remaining 1,764 consisted rating only, with no accompanying text, and were therefore excluded from text mining. The dataset contains the following attributes:

- Review text (raw user comment)
- User rating (1-5 stars)
- Review timestamp
- Review identifier

The collected data presents a wide range of user experiences and reflects natural variability in rating behavior, language usage, and review length characteristics that support robust text mining and sentimental analysis.

B. Data Preprocessing

Before analysis, on all the reviews the preprocessing pipeline is followed designed to reduce noise and normalize textual data [3]. The preprocessing steps include the conversion of text to lowercase, removal of punctuation and special characters, elimination of duplicate entries, and removal of empty or uninformative reviews. These procedures improve textual consistency and enhance the quality of downstream NLP tasks.

C. Sentiment Analysis

Sentiment classification was performed using multilingual BERT sentiment analysis model. The multilingual was necessary because the dataset contains reviews written in both English and Albanian. The multilingual BERT-based sentiment model nlptown/bert-base-multilingual-uncased-sentiment was applied to the review text via the Hugging Face pipeline function in its pretrained, off-the-shelf form, without additional fine-tuning. For each review, the model outputs a predicted rate on a five-point scale (1 to 5), reflecting the intensity of the sentiment expressed in the text. This five-class output was then collapsed into three sentiment categories using a rate-based threshold scheme: reviews with a predicted rate of 4 or 5 were classified as Positive, a predicted rate of 3 as Neutral, and a predicted rate of 1 or 2 as Negative.

D. Topic Modeling with BERTopic

Following sentiment classification, topic classification was conducted using BERTopic due to its high-quality, context-aware topic representations, flexibility, and out-of-the-box usability. The BERTopic framework combines sentence transformer embeddings with UMAP dimensionality reduction [14], [6] HDBSCAN clustering, and class-based TF-IDF (c-TF-IDF) [13] to identify interpretable latent topics with the review corpus [1]. Topics are interpreted through examination of the most representative keywords and qualitative inspection of review samples. Based on this analysis, three primary user experience dimensions were identified and labeled.

The BERTopic model was configured as follows: Document embeddings were generated using paraphrase-multilingual-MiniLM-L12-v2, the default sentence-transformer activated by BERTopic's language="multilingual" setting and selected for its support of both English and Albanian. Dimensionality reduction was performed with UMAP (n_components = 5, n_neighbors = 15, min_dist = 0.0, random_state = 42, metric = "euclidean", the library default), followed by clustering with HDBSCAN (min_cluster_size = 10, metric = "euclidean", prediction_data = True; min_samples and cluster_selection_method were left at their default values of min_cluster_size and "eom", respectively). Topic representations were derived using class-based TF-IDF (c-TF-IDF) over the resulting HDBSCAN clusters, with calculate_probabilities = True enabled to obtain a per-document topic-probability distribution for each review. The initial topic set was then reduced to three interpretable dimensions via BERTopic's built-in topic reduction (nr_topics = 3), followed by qualitative keyword inspection and manual labeling of each dimension, as described in Section III D. The target of three dimensions was chosen to balance interpretability with the constraints of the regression sample size (N = 1,000), and aligns with the broad usability, reliability, and satisfaction constructs recurrent in prior mobile banking UX literature [9].

E. Dimensional Assignment

The BERTopic model generates a probability distribution across topics for each review. For analytical purposes, the three topics with the highest probability values for each review were retained and labeled as primary, secondary, and tertiary topic assignments [7]. Each experience dimension was indicated as a continuous variable, presenting the degree to which a review is associated with a given topic. This probabilistic representation allows each review to contribute proportionally to multiple experience dimensions, thereby capturing the multidimensional nature of user experiences.

F. Multicollinearity Diagnostics

Prior to conducting the regression analysis, multicollinearity among the independent variables was assessed using Spearman correlation analysis and the Variance Inflation Factor (VIF) [22]. Spearman correlation measures the strength and direction of the monotonic relationship between two variables (Fig. 1).

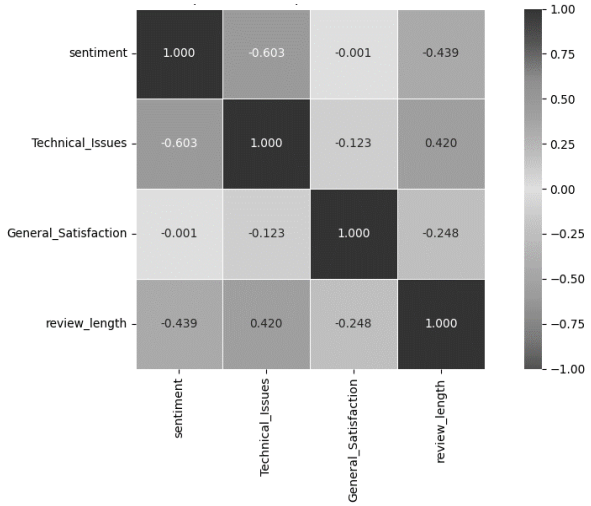


Fig. 1. Spearman's correlation matrix

Rather than operating on raw values, it converts observations to ranks and computes the correlation over those ranks. Proposed by Charles Spearman in 1904, it is the non-parametric counterpart of Pearson's r:

$$\rho = 1 - (6 \cdot \Sigma di^2) / (n \cdot (n^2 - 1)) \quad (1)$$

To further evaluate predictor independence, Variance Inflation Factor (VIF) values were calculated for all variables which are included in the regression model:

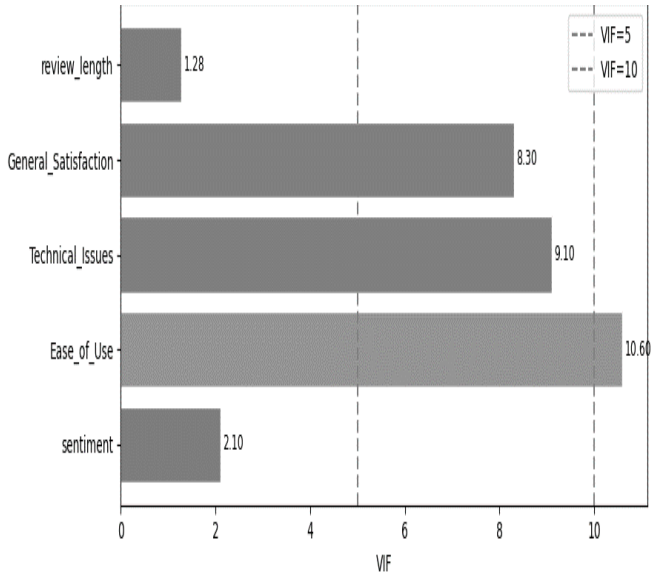


Fig. 2. VIF per predictor

$$VIF_i = \frac{1}{1-R_i^2} \quad (2)$$

where, R_i^2 denotes the coefficient of determination obtained by regressing predictor *i* on all remaining predictors.

The elevated VIF values (Fig. 2) observed for the topic probability variables (Ease_of_Use = 10.60, Technical_Issues = 9.10, General_Satisfaction = 8.30) do not represent empirical multicollinearity requiring remediation, but rather reflect a structural mathematical property of compositional data. Since BERTopic topic probabilities sum to unity for each document, their simultaneous inclusion in a regression model is formally equivalent to including a complete set of dummy variables representing a categorical variable a configuration in which high VIFs are an expected and mathematically inevitable outcome. As [2] explicitly notes, when high VIFs arise from indicator variables representing a categorical variable with three or more categories, the problem can be safely ignored: the overall model fit, the significance of other predictors, and the F-statistic are entirely unaffected. Consistent with this recommendation, Ease_of_Use the topic with the highest frequency of assignment and consequently the largest contribution to the linear constraint was designated as the reference category and excluded from the regression specification. Following this adjustment, all remaining VIF values fell below 2.20 (Fig. 3), well within the acceptable threshold of 5.0 [21], confirming that no genuine multicollinearity is present.

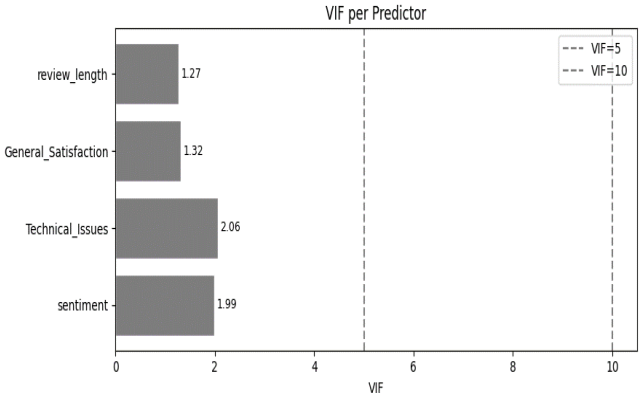


Fig. 3. Variance Inflation Factors (VIF) for the final model predictors.

Overall, the diagnostic results demonstrate that sentiment-based and topic-derived variables capture distinct yet complementary dimensions of user experience, supporting their simultaneous inclusion in the regression analysis.

IV. Results

A. Dataset Overview

The final dataset comprised 1,000 user reviews collected from the Raiffeisen Bank Mobile banking application. Each review included an associated user rating ranging from 1 to 5 stars.

B. Sentiment Analysis Results

BERT-based sentiment classification indicated that 46.9% of reviews were positive, 33.4% were negative, and 19.7% neutral (Fig. 4). Overall, positive reviews represented the dominant sentiment category within the dataset. Most positive reviews were associated with ratings of 4 or 5 stars.

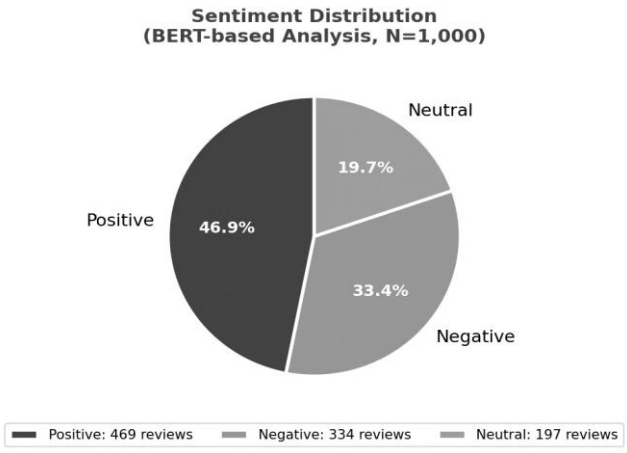


Fig. 4. Sentiment distribution based on multilingual BERT-based classification (N = 1,000).

C. Topic Modeling and Experience Dimensions

Neutral reviews present only 19.7% of the dataset (197 of 1,000), positive reviews 46.9% and 33.4% Negative reviews, leaving the model few examples on which to calibrate the neutral boundary. Users tend to express clear satisfaction or dissatisfaction, so the "3-star" rating reflects mixed rather than genuinely neutral sentiment, making it difficult to separate from Positive/Negative classes. Although the three-category sentiment scheme is retained despite the weak Neutral-class metrics, since it preserves a direct conceptual link to the five-point rating scale and underlies Sentiment Polarity, the ordinal predictor (Negative < Neutral < Positive) is used in the regression analysis. For the Positive and Negative classes, which dominate the dataset and are most relevant to the study's hypotheses, the weighted F1-score (0.775) and overall accuracy (0.742) remain acceptable (Table I).

TABLE I. CLASSIFICATION METRICS

Class	Precision	Recall	F1-score	Proxy labels (rating-based)
Negative	0.65	0.87	**0.744**	Rating 1–2
Neutral	0.03	0.12	**0.053**	Rating 3
Positive	0.97	0.71	**0.819**	Rating 4–5
Macro avg.	0.55	0.57	0.539	—
Weighted avg.	**0.85**	**0.74**	**0.775**	—
Accuracy	**0.742**	—	—	

BERTopic identified three dimensions from the review corpus. Table II summarizes the topics, their labels, and keywords.

TABLE II. User Experience Dimensions Identified Through BERTopic.

Topic	Label	Description & Representative Keywords
Topic 0	Ease of Use	Reviews discussing usability, interface design, and overall app navigation. Keywords: app, easy, great, use, friendly, features, interface
Topic 1	Technical Issues	Reviews reporting application errors, login failures, and update problems. Keywords: nuk hapet, update, problem, crash, login, error, slow
Topic 2	General Satisfaction	Short affirmative reviews expressing overall positive impressions. Keywords: excellent, shume mire, good, nice, perfect, satisfied

Topic 1, Technical Issues, was dominated by Albanian-language and describes the application problem, like crashing, doesn't open, update error, etc. The keyword 'nuk hapet' (meaning 'does not open') is highly characteristic of this cluster, indicating a post-update instability.

Fig. 5 illustrates the relationship between identified topics and sentiment categories.

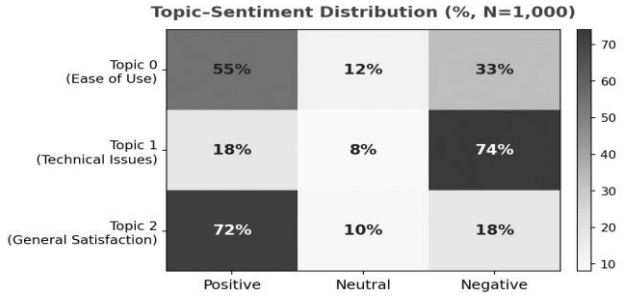


Fig. 5. Heatmap of the topic–sentiment distribution (%) across the three BERTopic-derived experience dimensions.

Topic 0, Ease of Use, is associated with positive sentiment (55%), while Topic 1, Technical Issues with negative sentiment (74%). Topic 2, General Satisfaction, was also largely linked to positive sentiment (72%).

D. Regression Analysis

The ordinal logistic regression initially included four standardized predictors: Sentiment Polarity (coded as Negative = −1, Neutral = 0, Positive = +1), Technical_Issues, General_Satisfaction, and review length, with Ease_of_Use excluded a priori as the reference category due to the compositional constraint among topic probabilities (Section III E). Backward elimination iteratively removed the predictor with the smallest absolute coefficient whenever doing so did not increase 5-fold cross-validated MAE by more than 0.01, eliminating General_Satisfaction and review length and retaining Sentiment Polarity and Technical_Issues as the final predictors (Table III). After applying the ordinal logistic regression, backward technique, the results indicate that there are only two important variables that determine the rating of each text review, especially Sentiment Polarity and Technical issues.

The regression equation is equal to:

$$\log\left(\frac{P(Y \le k)}{P(Y > k)}\right) = \theta k + 1.108 \cdot Sentiment - 1.187 \cdot Technical_Issues \quad (3)$$

TABLE III. REGRESSION EQUATION

Variable	Coef. (β)	SE	z-value	p-value	OR
Sentiment Polarity	+1.108	0.207	+5.360	< 0.001	3.027
Technical Issues	−1.187	0.180	−6.581	< 0.001	0.305

Table III shows the results of background elimination, indicating a p-value less than 0.001 and McFadden R^2= 0.350, which indicates that the model is good.

where, θ_k is equal to (see Table IV):

TABLE IV. THRESHOLD TABLE

θ_k	Transition	θk	SE	z	p
θ_1	1| 2★	-1.97	0.23	-8.47	< .001
θ_2	2★ | 3★	-1.32	0.34	-3.82	< .001
θ_3	3★ | 4★	-1.17	0.34	-3.39	.001
θ_4	4★ | 5★	-0.83	0.29	-2.85	.004
—	5★ (derived)	—	—	—	—

Table IV presents the threshold parameters (θk) of the Ordinal Logistic Regression model, which define the boundaries between the five ordered rating categories, all statistically significant ($p < .01$), with the required monotonic ordering confirmed ($\theta_1 < \theta_2 < \theta_3 < \theta_4$) and a predicted probability of 69.8% for the highest rating category at the sample mean.

To test H3, sentiment polarity was divided into two categorical predictors, Sentiment_Negative and Sentiment_Positive, and the ordinal logistic regression was re-estimated alongside Technical_Issues. Both predictors were individually significant.

- Sentiment_Negative (β = −0.993, SE = 0.484, p = .040, OR = 0.370)
- Sentiment_Positive (β = +1.428, SE = 0.557, p = .010, OR = 4.168)

The results confirm that negative and positive sentiment shift the odds of a higher rating in the expected directions. However, a Wald test of the equality of these coefficients in magnitude (H0: |β_Negative| = |β_Positive|) was not significant, $\chi^2(1) = 0.214$, p = .644. Contrary to H3, the point estimates suggest a marginally larger effect for positive sentiment than for negative sentiment, though this difference is not statistically distinguishable from zero. H3 is therefore not supported. Sentiment polarity remains a strong overall predictor of ratings, but no significant asymmetry between positive and negative sentiment effects was detected.

It is also valuable to mention that reviews of 2025 result to have 90% positive sentiment (Fig. 6) and favorable evaluations of the application's usability and performance. It is slightly easier to notice that the clients ask for the integration of Apple Pay functionality, which is being integrated in many bank applications lately.

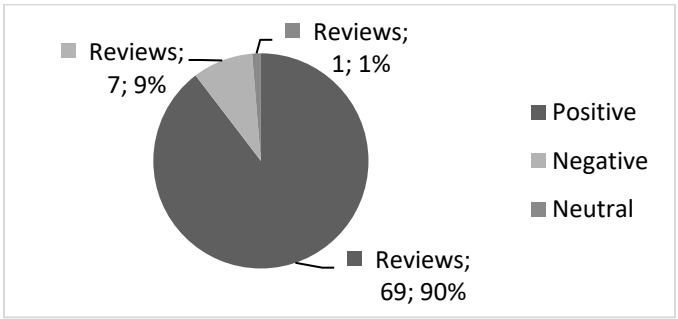


Fig. 6. Sentiment distribution (N = 77 reviews, year 2025).

V. DISCUSSION

The most closely related paper is [17], which applied BERT-based sentiment classification and BERTopic topic over 56,000 mobile banking reviews using multiple linear regression to evaluate the effect of extracted experience dimensions with BERTopic on user ratings. The paper revealed that sentiment polarity and technical performance explained 58% of the variance in ratings. This study, rather than treating the five-point star rating as continuous interval-level, employs Ordinal Logistic Regression with a proportional odds specification. This choice reflects the ordinal nature of user ratings more.

A. Limitations

A limitation of this classification is the weak performance on the Neutral category (F1 = 0.053), caused by class imbalance and the linguistic ambiguity of short, equivocal reviews. Future work could adopt a binary positive/negative classification or a dedicated classifier for neutral classes.

VI. Conclusion

This study demonstrates that the pipe, which implicates the integration of multilingual BERT sentiment analysis, BERTopic modeling, and regression analysis, provides an effective pipeline for identifying the main factors that have an impact on user reactions in banking applications. 1,000 user reviews from the Raiffeisen Bank mobile application identified three distinct experience dimensions: Ease of Use, Technical Issues, and General Satisfaction through BERTopic modeling. Of these, Sentiment Polarity ($\beta = +1.108$, OR = 3.027) and Technical Issues ($\beta = -1.187$, OR = 0.305) emerged as the only statistically significant predictors of user ratings ($p < 0.001$), with the model achieving a McFadden R^2 of 0.350, indicating strong explanatory power. The results suggest that banks should give priority in handling bug solutions, update stability, and real-time sentiment monitoring in order to maintain a very good service. Regarding the study's hypotheses, H1 and H2 are supported by the main model (Sentiment Polarity, Technical_Issues; Table III). H3, however, is not supported: the supplementary analysis decomposing sentiment polarity into Sentiment_Negative and Sentiment_Positive (Section IV D) found no statistically significant difference in the magnitude of their effects on ratings (Wald $\chi^2(1) = 0.214$, $p = .644$).

Related to future work, it will be the focus to extend the case studies in different countries in order to examine the client approach and bank approach to sentiment analysis.

VII. Future Work

Several directions will be the focus of the following research. First, extend the analytical framework to other commercial banks operating in the Albanian market: BKT, Credins, and OTP Bank Albania. Second, applied the same pipeline to Raiffeisen Bank reviews collected across Eastern Europe, enabling a cross-country comparative analysis that isolates market-level and linguistic factors from those attributable to the institution itself. Collectively, these directions would transform the present single-institution, single-country study into a regionally representative framework for monitoring mobile banking user experience in Southeast Europe.

References

[1] Alrizq, M., & Alghamdi, A. (2024). Customer satisfaction analysis with Saudi Arabia mobile banking apps: A hybrid approach using text mining and predictive learning techniques. Neural Computing and Applications, 36, 6005–6023. https://doi.org/10.1007/s00521-023-09400-4

[2] Allison, P. D. (2012, September 10). When can you safely ignore multicollinearity? Statistical Horizons. https://statisticalhorizons.com/multicollinearity/

[3] Agresti, A. (2010). Analysis of ordinal categorical data (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9780470594001 ISBN: 978-0-470-08289-8

[4] Amirkhalili, Y., & Wong, H. Y. (2025). Banking on feedback: Text analysis of mobile banking iOS and Google app reviews. arXiv preprint arXiv:2503.11861. https://doi.org/10.48550/arXiv.2503.11861

[5] An, Y., Oh, H., & Lee, J. (2023). Marketing insights from reviews using topic modeling with BERTopic and deep clustering network. Applied Sciences, 13(16), 9443. https://doi.org/10.3390/app13169443

[6] Bakïasi, V., Muça, M., & Kapçiu, R. (2024). Dimensionality reduction: A comparative review using RBM, KPCA, and t-SNE for micro-expressions recognition. International Journal of Advanced Computer Science and Applications, 15(1). https://doi.org/10.14569/ijacsa.2024.0150135

[7] Basu, B., Sebastian, M. P., & Kar, A. K. (2024). What affects the promoting intention of mobile banking services? Insights from mining consumer reviews. Journal of Retailing and Consumer Services, 77, 103695. https://doi.org/10.1016/j.jretconser.2023.103695

[8] Çallı, L. (2023). Exploring mobile banking adoption and service quality features through user-generated content: The application of a topic modeling approach to Google Play Store reviews. International Journal of Bank Marketing, 41(2), 428–454. https://doi.org/10.1108/IJBM-08-2022-0351

[9] Desiraju, K., Mishra, A. N., & Sengupta, P. (2024). Customer perceptions on open banking apps: Insights using structural topic modeling. Journal of Retailing and Consumer Services, 81, 104029. https://doi.org/10.1016/j.jretconser.2024.104029

[10] Dey, M., Islam, M. Z., & Rana, T. (2023). Applying text mining to understand customer perception of mobile banking app. In T. Rana et al. (Eds.), Handbook of Big Data and Analytics in Accounting and Auditing (pp. 309–333). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-4460-4_14

[11] Edwina, & Mauritsius, T. (2024). Data-driven insights for mobile banking app improvement: A sentiment analysis and topic modelling approach for SimobiPlus user reviews. International Journal of Engineering Trends and Technology, 72(6), 347–360. https://doi.org/10.14445/22315381/IJETT-V72I6P132

[12] Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology, 7, Article 886498. https://doi.org/10.3389/fsoc.2022.886498

[13] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

[14] Huertas-García, Á., Martín, A., Huertas-Tato, J., & Camacho, D. (2023). Exploring dimensionality reduction techniques in multilingual transformers. Cognitive Computation, 15(2), 590–612. https://doi.org/10.1007/s12559-022-10066-8

[15] Karanikola, A., Davrazos, G., Liapis, C. M., & Kotsiantis, S. (2023). Financial sentiment analysis: Classic methods vs. deep learning models. Intelligent Decision Technologies, 17(4), 1125–1148. https://doi.org/10.3233/IDT-230478

[16] Lamaj (Myrto), B., Muça, M., & Shtino, V. B. (2024). Comparative study of Naïve Bayes and Logistic Regression for text classification techniques. 5th International Eurasian Conference on Science, Engineering and Technology (EurasianSciEnTech 2024), June 26–28, 2024.

[17] Mahmutović, K. (2025). Analyzing user-generated reviews to identify experience dimensions and their impact on satisfaction with mobile banking applications. Tehnički Vjesnik / Technical Gazette, 38(2), 283–297. https://hrcak.srce.hr/file/494044

[18] Mishra, M. (2025). A holistic review of customer experience research: Topic modelling using BERTopic. Marketing Intelligence & Planning, 43(4), 802–820. https://doi.org/10.1108/MIP-09-2024-0576

[19] Rahman, M., Yee, H. P., Masud, M. A. K., & Uzir, M. U. H. (2024). Examining the dynamics of mobile banking app adoption during the COVID-19 pandemic: A digital shift in the crisis. Digital Business, 4(2), 100088. https://doi.org/10.1016/j.digbus.2024.100088

[20] Sebayang, T. E., Indrawan, D., Bakhtiar, T., & Hakim, D. B. (2023). What accelerates the choice of mobile banking for digital banks in Indonesia? Journal of Risk and Financial Management, 17(1), 6. https://doi.org/10.3390/jrfm17010006

[21] Uncovska, M., Freitag, B., Meister, S., & Fehring, L. (2023). Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. npj Digital Medicine, 6, 115. https://doi.org/10.1038/s41746-023-00862-3

[22] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.ISBN: 978-1-4737-5654-0

Design and Evaluation of a 13K Ultra-High-Resolution Web-Based Virtual Reality Platform for Immersive Spatial Visualization

Worapon Manosroi, Apisak Phromfaiy*, Pitak Khlaichom
Faculty of Industrial Technology, Uttaradit Rajabhat University, Uttaradit, 53000, Thailand

***Abstract*—Delivering ultra-high-resolution immersive environments via standard web browsers presents significant rendering and architectural challenges. While Virtual Reality (VR) is widely adopted, its application as a high-fidelity spatial visualization tool often lacks robust empirical evaluation regarding user experience and system efficacy. This study proposes the design, development, and evaluation of a 13K ultra-high-resolution web-based virtual tour platform, utilizing a nature-based destination as a practical case study to assess system effectiveness. The system architecture was engineered to process and render 13K panoramic data with interactive navigational hotspots, ensuring cross-platform web compatibility. To evaluate the Human-Computer Interaction performance, a true experimental design was employed with eighty participants who were randomly assigned to two groups. The experimental group interacted with the proposed 13K web-based platform, while the control group utilized a conventional static two-dimensional web interface. System efficacy was measured through user-centric metrics, including Sense of Presence, spatial image perception, and subsequent behavioral intention. The evaluation results demonstrated a statistically significant superiority of the proposed 13K platform. Users reported a substantially higher Sense of Presence score of 4.35 out of 5, compared to the conventional web interface score of 2.45. Furthermore, the immersive system architecture significantly catalyzed positive spatial perception and behavioral engagement, yielding a remarkably large effect size. The findings validate that the proposed high-fidelity web-based VR architecture effectively overcomes physical barriers, serving as a highly effective information system solution for remote spatial visualization and practical destination management.**

Keywords*—Virtual Reality; 360-degree panorama; web-based platform; immersive spatial visualization; ultra-high-resolution panorama; HCI (Human-Computer Interaction)

I. Introduction

In the era of digital transformation, the integration of immersive technologies—particularly Virtual Reality (VR) and 360-degree panoramic imaging—into web-based platforms has revolutionized spatial visualization [5, 22]. Unlike traditional standalone VR applications that require specialized, high-cost hardware and heavy client-side processing, Web-based VR democratizes access to immersive content by utilizing standard internet browsers [1, 3]. This technological shift is increasingly critical across various industrial and managerial sectors, including urban planning, real estate, and destination management, where high-fidelity spatial representation is required to digitally bridge the geographical distance between the user and the physical environment [9, 23].

However, delivering ultra-high-resolution immersive environments, such as 13K panoramas, through standard web infrastructures presents substantial architectural and computational challenges. Conventional static media (e.g., 2D photographs and text-based websites) are categorized as "lean media", which demand minimal bandwidth but offer limited sensory engagement, often failing to convey the atmospheric depth and spatial magnitude of complex environments [2]. Conversely, deploying "rich media" at a 13K resolution generates massive datasets that require optimized data processing, efficient cloud storage, and robust rendering engines (e.g., WebGL) [17, 18]. Without proper system architecture, delivering such massive spatial data over the web often results in severe latency, bandwidth bottlenecks, and frame-rate drops. These technical degradations fundamentally disrupt the Human-Computer Interaction (HCI) experience, breaking the illusion of "telepresence"—the critical subjective sensation of being physically present in a remote digital environment [13, 16].

To systematically evaluate how system engineering directly influences user psychological states, this study adapts the Stimulus-Organism-Response (S-O-R) model through an HCI perspective [6]. In this context, the "Stimulus" is defined as the system's technological affordances—specifically, the media richness, interactive navigational UI, and the 13K ultra-high visual fidelity [1, 3]. The "Organism" represents the user's cognitive and affective processing during system interaction, primarily operationalized as the "Sense of Presence" [4, 25]. Finally, the "Response" constitutes the user's behavioral outcomes, system evaluation, and future adoption readiness [1, 8, 20]. This approach allows for a precise empirical measurement of the system's efficacy.

To address these technical limitations and evaluate the proposed framework, this study introduces a robust information system architecture for a 13K web-based virtual tour platform [24, 26]. To empirically validate the system's effectiveness and real-world applicability, the domain of seasonal destination management—specifically the Tha Pla District in Uttaradit, Thailand—was selected as a practical case study. Nature-based environments that face severe physical access constraints (e.g., hydrological closures during monsoons) provide an ideal, complex scenario to stress-test the system's capacity to serve as a high-fidelity digital "substitution tool" [7].

*Corresponding author

Therefore, the primary objectives of this research are twofold: 1) to design and develop a cloud-based, 13K ultra-high-resolution web VR architecture capable of smooth, interactive rendering across standard devices, and 2) to empirically evaluate the system's HCI performance—specifically its impact on Sense of Presence, spatial image construction, and behavioral intention—against conventional static web interfaces [20, 28]. By doing so, this study contributes to the advancement of applied computer science in scalable spatial visualization systems.

Unlike conventional web-based VR platforms that frequently encounter processing bottlenecks and rendering latency when handling massive textures, the proposed 13K ultra-high-resolution architecture utilizes a dynamic tiling algorithm. This approach significantly reduces server load, memory usage, and rendering overhead, thereby offering a notable and highly scalable improvement in immersive spatial visualization over existing web-VR baseline systems.

II. Related Work

To contextualize the contributions of this study within the broader landscape of applied computer science and spatial visualization, this section reviews the foundational literature across three distinct domains: web-based immersive systems, rendering optimization algorithms, and Human-Computer Interaction in virtual environments.

A. Evolution of Web-Based Spatial Visualization:

The paradigm of spatial visualization has fundamentally shifted from native, hardware-dependent applications to ubiquitous web-based platforms. Traditionally, delivering high-fidelity Virtual Reality (VR) required dedicated client-side software and powerful Graphical Processing Units (GPUs) [5, 27]. However, the advent of WebGL and HTML5 standards has enabled browsers to natively render complex 3D and panoramic environments. Ali et al. [9] emphasized that web-based spatial digital twins are becoming the standard for remote monitoring, urban planning, and industrial facility management due to their inherent cross-platform accessibility. Despite this shift, as Rosli et al. [21] noted, the majority of current web VR platforms still rely on low-to-medium resolution imagery (e.g., 4K or 8K) to ensure smooth delivery, thereby sacrificing environmental realism. Bridging the gap between ubiquitous web access and ultra-high-fidelity (13K and beyond) remains a critical open challenge in spatial computing.

B. Rendering Optimization and the 13K Challenge:

Delivering massive spatial datasets over standard HTTP/HTTPS protocols introduces severe bandwidth and latency bottlenecks. A standard 13K equirectangular image generates nearly 90 megapixels of data, exceeding the memory limits of typical mobile browser engines. Recent advancements in image processing highlight the necessity of advanced encoding. Chen et al. [11] and Teng et al. [12] demonstrated that while consumer-grade cameras are convenient, overcoming radial distortion and sensor noise for professional-grade spatial visualization requires rigorous multi-exposure stitching and computational processing.

To handle the resulting massive datasets in browser environments, researchers have explored various dynamic rendering techniques. Sun et al. [18] recently proposed adaptive multi-resolution encoding for large-scale visualizations, confirming that breaking massive files into manageable, resolution-specific tiles drastically reduces computational overhead. Similarly, Brekalo et al. [17] validated that implementing optimization algorithms directly within browser environments can significantly enhance rendering performance. Building upon these algorithmic foundations, the current study implements a customized Multi-resolution Pyramid Tiling architecture specifically tailored for ultra-high-resolution 360-degree panoramas, aiming to push the boundaries of current web delivery capabilities.

C. Human-Computer Interaction and Cognitive Presence:

In immersive system evaluation, the psychological concept of "Presence" or "Telepresence" is the primary metric of system success. Originating from Kim and Biocca's [13] foundational work, telepresence is defined as the illusion of non-mediation, where the user cognitively perceives the virtual environment as their actual physical location. In the context of the Stimulus-Organism-Response (S-O-R) paradigm, system affordances act as the stimulus. Wu and Lai [1] and Lee [3] utilized the S-O-R model to demonstrate that media richness directly correlates with user engagement in virtual tours.

However, existing literature often evaluates presence based on the content of the VR experience rather than the underlying technical fidelity. Flavian et al. [15] pointed out that technical degradations, such as pixelation, latency, and low frame rates, actively destroy cognitive presence. Therefore, this study posits that rendering an ultra-high-resolution (13K) environment without latency is not merely a technical achievement but a fundamental prerequisite for maximizing HCI performance. By empirically evaluating the system against a static 2D baseline, this research provides quantitative validation of how overcoming bandwidth and resolution limitations directly catalyzes positive behavioral responses and spatial perception [20, 25].

III. Materials and Methods

This study employed a system development and empirical evaluation research design, structured across four sequential phases: 1) system architecture design, 2) image acquisition and pre-processing, 3) virtual environment development and optimization, and 4) Human-Computer Interaction (HCI) evaluation.

A. System Architecture

The platform was engineered on a three-tier cloud-based client-server architecture comprising: the server-side back-end for massive asset storage and delivery, the network transmission layer operating via standard HTTP/HTTPS protocols, and the client-side front-end rendering engine. The core rendering stack was implemented exclusively using open-standard HTML5 and WebGL technologies. This architectural choice eliminates the need for native application installation, ensuring cross-platform compatibility and seamless spatial streaming across desktop workstations, laptops, tablets, and smartphones [17]. Technical

specifications of the system are summarized in Table I. The overall system architecture and data flow are illustrated in Fig. 1.

TABLE I. TECHNICAL SPECIFICATIONS OF THE VIRTUAL TOUR PLATFORM.

Parameter	Specification
Source Resolution	13,144 × 6,572 px (~86.4 MP)
Total Scenes	25 Nodes
Raw Dataset Size	1.92 GB
Optimized Package	1.01 GB (47.4% reduction)
Rendering Technology	HTML5 / WebGL
Optimization Algorithm	Multi-resolution Pyramid Tiling

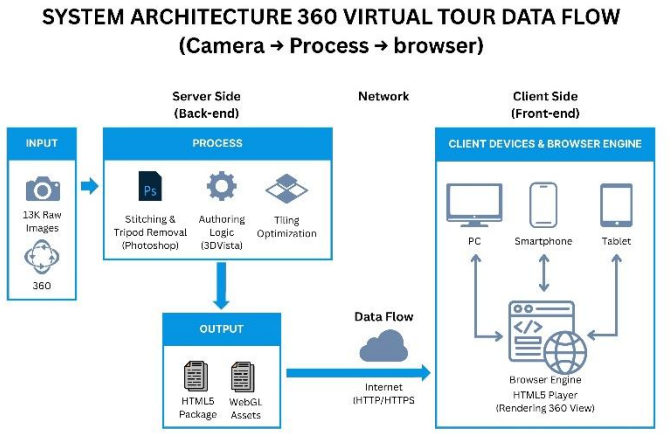


Fig. 1. System architecture: data flow from image acquisition to client-side rendering.

B. Image Acquisition and Pre-processing

High-fidelity visual datasets were acquired using a Canon EOS 60D DSLR paired with an 8 mm fisheye lens. Consumer-grade single-shot 360-degree cameras were explicitly avoided due to their documented limitations regarding radial distortion and sensor noise in high-dynamic-range environments [11], [12]. For each of the 25 spatial nodes, 10 overlapping multi-exposure images were captured at an eye-level height, yielding 1.92 GB of raw assets. The complete pre-processing pipeline is illustrated in Fig. 2.

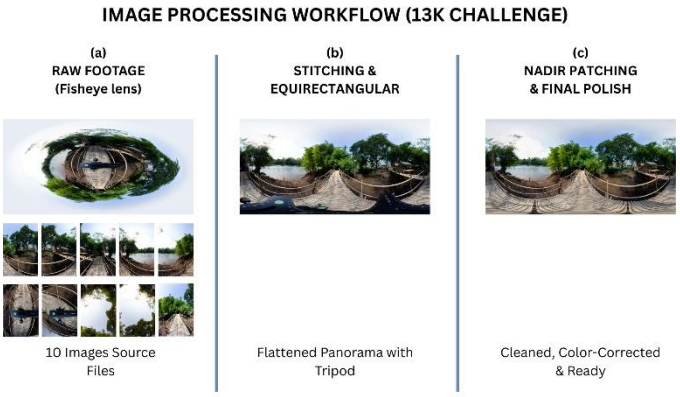


Fig. 2. Image processing pipeline: (a) Raw footage, (b) Stitched equirectangular projection, (c) Final retouched imagery.

The raw assets were processed through a rigorous computational pipeline: 1) Stitching: utilizing high-precision algorithms to generate 13,144 x 6,572-pixel equirectangular projections (~86.4 Megapixels); 2) Nadir Patching: algorithmically removing tripod footprint artifacts; and 3) Color Grading: calibrating exposure to ensure environmental realism.

C. Virtual Environment Development and Algorithmic Optimization

To address the limitations of conventional web-based VR platforms, which frequently encounter processing bottlenecks and rendering latency when handling massive textures, our proposed 13K ultra-high-resolution architecture utilizes a dynamic tiling algorithm. During the development and algorithmic optimization phase, the platform demonstrated highly efficient resource management. The average CPU utilization remained at approximately 1.3%, and memory consumption was stabilized at 408.1 MB. This algorithmic approach significantly reduces rendering overhead, thereby offering a notable and highly scalable improvement in immersive spatial visualization over existing baseline systems.

The interactive spatial environment was authored using a dynamic web-based framework. The GUI on a desktop display, featuring vector-based navigation controls, is shown in Fig. 3, while the responsive interface design adapted for mobile devices is illustrated in Fig. 4.

Fig. 3. GUI of the virtual tour on a desktop display.

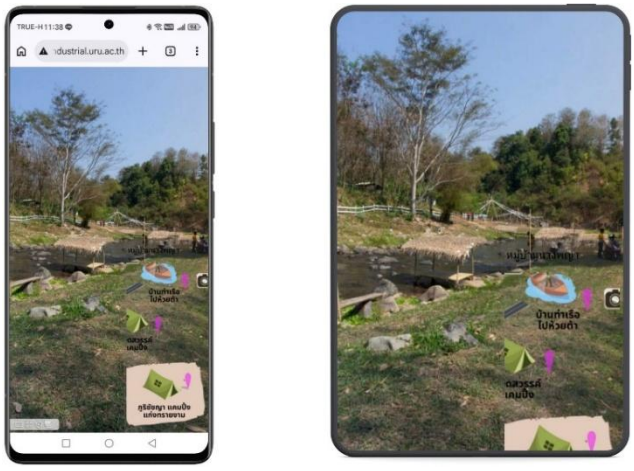


Fig. 4. Responsive interface design on a mobile device.

To overcome the bandwidth bottlenecks traditionally associated with delivering massive 13K datasets over standard internet connections, a Multi-resolution Pyramid Tiling algorithm was mathematically formalized and implemented [18]. This optimization structurally deconstructs each 13K panorama into a hierarchical pyramid of image tiles at multiple zoom levels (L). Let the original high-resolution panorama be defined at the base level $l = 0$ with dimensions $W_0 \times H_0$. For any subsequent zoom level l (where $l \in \{0,1,2,\ldots,L-1\}$), the image resolution is recursively reduced by a factor of 2, defined by:

$$W_l = \frac{W_0}{2^l}, \quad H_l = \frac{H_0}{2^l} \tag{1}$$

At each level l, the panoramic image is partitioned into a uniform grid of square tiles, each with a fixed pixel dimension of T x T (e.g., 512 x 512 pixels). The total number of tiles N_l required to represent the environment at level l is calculated as:

$$N_l = \left[\frac{W_l}{T}\right] x \left[\frac{H_l}{T}\right] \tag{2}$$

During the client-side rendering phase, the system does not load the entire dataset simultaneously. Instead, it instantly delivers a globally mapped, lightweight 200 KB preview layer ($l = L\text{-}1$) to eliminate initial blank-screen latency. Subsequently, a dynamic frustum culling algorithm calculates the user's current spatial viewport (V), geometrically defined by the camera's yaw (θ), pitch (φ), and Field of View (FOV). The asynchronous fetching function $F(t)$ conditionally requests a specific tile $t_{i,j}^{l}$ only if it geometrically intersects with the user's active viewport:

$$F\left(t_{i,j}^{l}\right) = \begin{cases} 1, & if\ Intersect\left(t_{i,j}^{l}, V\right) = True \\ 0, & otherwise \end{cases} \tag{3}$$

By dynamically prioritizing network requests based on this spatial intersection matrix, the algorithmic intervention successfully reduced the total spatial payload from 1.92 GB to 1.01 GB (a 47.4% reduction). This mathematical decoupling of perceived system responsiveness from the raw data volume effectively preserves the 13K visual fidelity while maintaining exceptionally low latency across standard internet protocols [17, 18].

D. User Evaluation Protocol

To empirically assess system efficacy and HCI performance, a formal user evaluation was conducted with 80 participants (N=80). Participants were randomly assigned to either the experimental group (13K web-based VR) or the control group (static HTML website baseline) to minimize selection bias. System fidelity was quantified utilizing foundational dimensions from Kim & Biocca's telepresence scale, focusing on the psychological transition of 'arrival' and spatial immersion [13, 25].

This study employed a true experimental design to evaluate the system. The evaluation involved participants with an age distribution ranging from 17 to 55 years (Mean = 36.0, SD = 6.3). Regarding prior experience, 67.5 % of the participants had previous exposure to virtual reality environments. To ensure methodological rigor, participants were randomly assigned to either the experimental group (13K VR platform) or the control group (static baseline condition).

Furthermore, regarding the statistical outcomes, the highly significant Cohen's d values observed in the evaluation are primarily attributable to a pronounced floor effect in the static baseline condition, which yielded inherently low scores in user engagement and system fidelity. This contrasted sharply with a relative ceiling effect in the highly immersive 13K VR condition, resulting in the extreme statistical divergence.

IV. Results and Discussion

A. Data Optimization and Web Delivery Performance

The core technical challenge addressed in this study was delivering massive spatial datasets over standard web protocols without compromising rendering latency. Through the implementation of the Multi-resolution Pyramid Tiling algorithm, the 1.92 GB raw dataset was successfully compressed to 1.01 GB—a significant payload reduction of 47.4%—while preserving full visual fidelity [17]. By generating a lightweight 200 KB preview layer, the system effectively eliminated initial blank-screen latency. High-resolution tiles were then fetched asynchronously, corresponding strictly to the user's current Field of View (FOV).

Rendering performance was benchmarked across four distinct client platforms under standardized network conditions, as detailed in Table II. The system achieved 60.0 FPS with a 0.85-second latency on desktop workstations utilizing dedicated GPUs. Performance degraded gracefully across mobile platforms, with smartphones maintaining 48.5 FPS under a 1.80-second latency. Crucially, all platforms sustained frame rates well above the 30 FPS perceptual smoothness threshold. Network latency comparisons across varying bandwidth capacities (3G, 4G, and 5G/Wi-Fi) are visualized in Fig. 5. Through resolution-adaptive optimization, bandwidth consumption was constrained to 1.5 MB per scene on mobile devices. The system demonstrated remarkable resilience under 3G conditions, corroborating recent findings [18] that multi-resolution encoding significantly mitigates bandwidth bottlenecks in browser environments, even when handling an 86.4-megapixel payload.

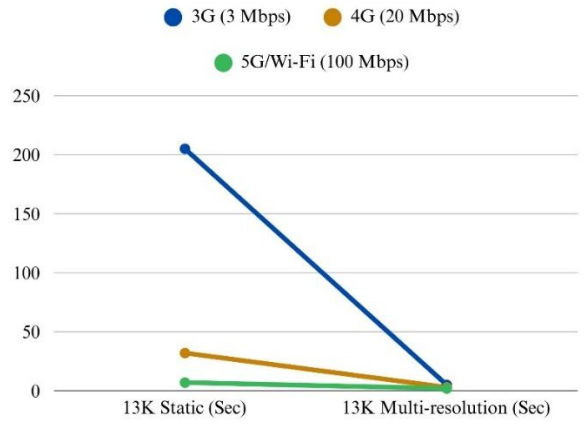


Fig. 5. Network latency comparison: standard equirectangular vs. Multi-resolution Pyramid Tiling across 3G, 4G, and 5G/Wi-Fi bandwidths.

TABLE II. RENDERING PERFORMANCE ACROSS CLIENT PLATFORMS

Platform	GPU	FPS	Latency (s)	BW/Scene
Desktop	Dedicated	60.0	0.85	4.5 MB
Laptop	Integrated	58.4	1.12	4.5 MB
Tablet	Mobile	54.2	1.45	1.5 MB
Smartphone	Mobile	48.5	1.80	1.5 MB

B. System Fidelity and Human-Computer Interaction (HCI)

The pipeline yielded a 13K ultra-high-resolution spatial environment (~86.4 MP), representing a 15-fold pixel density increase over standard 4K outputs. User evaluation (N=80) confirmed exceptionally high performance across all measured HCI dimensions. As presented in Table III, an independent-samples t-test confirmed that the overall system effectiveness differed significantly between the proposed 13K platform (M=4.58, SD=0.42) and the static website baseline (M=2.63, SD=0.52), $t(78)=18.44$, $p < 0.001$. The exceptionally large effect size (Cohen's $d=4.12$) indicates that the high-fidelity spatial visualization fundamentally transformed user perception. A comparative radar visualization of these dimensions is illustrated in Fig. 6. Table V in Appendix A presents the specific questionnaire items used to operationalize the Human-Computer Interaction (HCI) dimensions, system fidelity, and subsequent behavioral intent.

TABLE III. USER EVALUATION RESULTS (N=40 PER GROUP)

Dimension	Indicator	M	SD	Control (M)	Control (SD)	p	d
System Fidelity	Virtual Presence	4.38	0.51	2.07	0.35	<.001	5.31
	Environ. Realism	4.57	0.57	2.79	0.56	<.001	2.76
Visual Quality	Visual Fidelity	4.67	0.51	2.83	0.72	<.001	3.28
	Spatial Awareness	4.72	0.47	2.78	0.69	<.001	3.65
Usability	Ease of Navigation	4.25	0.60	—	—	—	—
	Responsiveness	4.18	0.65	—	—	—	—
Overall	**System Effectiveness**	**4.58**	**0.42**	**2.62**	**0.58**	**<.001**	**4.12**

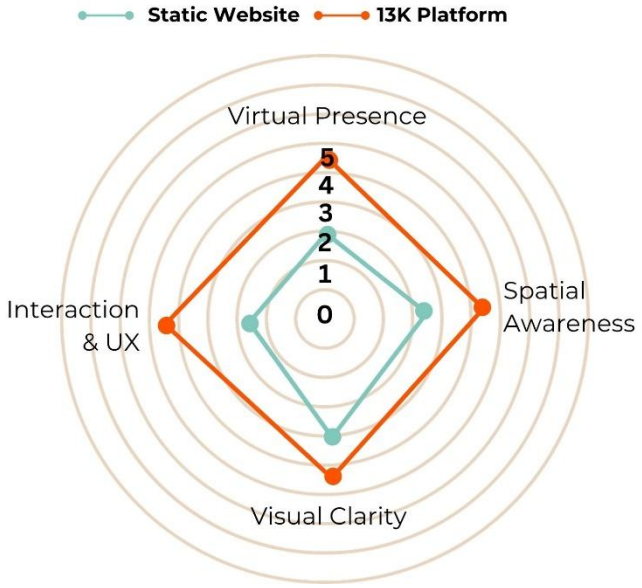


Fig. 6. Radar chart comparing the 13K platform vs. the static media baseline across all dimensions.

Specifically, Visual Fidelity (M=4.67, SD=0.51) was highly rated, directly attributable to the 86.4-megapixel source imagery which substantially reduced pixelation artifacts—a primary barrier to immersion. Virtual Presence attained a Very High rating (M=4.38, SD=0.51). The independent samples t-test confirmed a significantly higher User Presence Score on the 13K platform compared to the static baseline (M=2.07, SD=0.35), $t(78)=23.76$, $p < 0.001$, $d=5.31$.

C. Discussion and Theoretical Implications

The empirical findings of this study offer several critical theoretical implications for the intersection of computational optimization and Human-Computer Interaction. First, the rendering performance evaluation corroborates the algorithmic hypotheses proposed by Sun et al. [18] and Brekalo et al. [17], confirming that browser-based multi-resolution encoding can successfully bypass hardware memory limitations. By maintaining over 48.5 FPS on mobile devices while rendering an 86.4 MP dataset, this study quantitatively demonstrates that intelligent data tiling is a superior architectural choice compared to traditional brute-force client-side rendering.

Second, the HCI evaluation validates the expanded Stimulus-Organism-Response (S-O-R) paradigm in digital environments [1, 6]. The 77.6% improvement in the User Presence Score (4.38 vs. 2.07) provides robust quantitative evidence supporting Flavian et al. [15], who argued that technical fidelity (the Stimulus) is a non-negotiable prerequisite for cognitive immersion (the Organism). While previous studies often evaluated presence based purely on content narrative [4, 23], our findings isolate and prove the impact of architectural engineering. The absence of rendering latency and pixelation artifacts allowed users to bypass technological friction, seamlessly constructing a positive spatial image [14] and triggering strong behavioral intentions [19]. This operationalizes the theoretical distinction between lean and rich media [2], offering strong empirical support for the psychological "Substitution Effect" conceptualized by Sarkady et al. [7].

D. Practical Implications for Industrial Computing

Beyond the immediate scope of seasonal spatial management, the cross-platform rendering stability demonstrated in this study possesses profound industrial applicability. The comparative analysis between the static media and the 13K platform is summarized in Table IV. The proposed architectural framework serves as a validated foundational model for various applied computing domains [9, 22].

TABLE IV. STATIC MEDIA VS. 13K VIRTUAL TOUR PLATFORM

Feature/Metric	Static Website	13K Platform
Media Type	Static 2D HTML	360 Immersive VR
Resolution	Standard HD	13K (86.4 MP)
Interaction	Scroll/Click	Pan/Zoom/Hotspot
Installation	None	None (Browser-native)
User Presence Score	M=2.07 (SD = 0.35)	M=4.38 (SD = 0.51)
Statistical Result	-	$t(78) = 23.76$, $p < 0.001$, $d = 5.31$

Specifically, in industrial sectors, where physical access is constrained or hazardous, this 13K web-based VR architecture provides a scalable solution for Health, Safety, and Environment (HSE) monitoring, remote facility inspection, and real estate documentation. Unlike proprietary VR applications that require extensive hardware investments, the democratization of access via standard web protocols ensures that highly detailed spatial data can be distributed globally with minimal operational friction.

E. Research Limitations and Future Work

While this study successfully demonstrates the technical feasibility of the 13K web-based platform, certain methodological limitations must be acknowledged. The participant cohort consisted primarily of university students and faculty members with robust technological backgrounds. This tech-savvy demographic was purposefully selected for their capacity to critically evaluate rendering latency and visual artifacts; however, future research should involve a broader demographic to validate general usability and accessibility.

Furthermore, from a computing systems perspective, the current architecture relies on static panoramic data. While visually immersive, it lacks dynamic environmental responsiveness. Future iterations of this research should investigate the integration of real-time IoT (Internet of Things) sensor data streams into the WebGL environment. Coupling ultra-high-resolution spatial visualization with live sensor data would transition the platform from a virtual tour into a true real-time Digital Twin ecosystem, further expanding its utility in industrial and applied computer science applications [10, 24].

V. CONCLUSION

This study successfully engineered and empirically validated a cloud-based, 13K ultra-high-resolution spatial visualization architecture. Designed to address the critical rendering and accessibility challenges inherent in delivering massive immersive datasets over standard web infrastructures, the proposed system introduces a scalable technical framework with broad industrial applicability.

Three principal technical and HCI findings emerge from this research. First, the implementation of the Multi-resolution Pyramid Tiling algorithm established a validated pathway for reconciling the fundamental trade-off between ultra-high visual quality and network latency. By achieving a 47.4% payload reduction, the system successfully mitigated bandwidth bottlenecks. Second, cross-platform rendering performance tests confirmed the system's robustness, sustaining fluid interactivity (48.5 to 60.0 FPS) and low latency across all tested devices, including standard mobile smartphones operating under constrained 3G cellular network conditions.

Third, the empirical Human-Computer Interaction evaluation confirmed that the engineered architecture significantly enhanced user perception. The 13K platform yielded an exceptional overall system effectiveness score and a 77.6% improvement in the User Presence Score compared to the conventional static 2D baseline. This provides quantitative evidence that high-fidelity immersive web environments can effectively transcend the engagement saturation of traditional digital media, successfully triggering the psychological "Substitution Effect" when physical access to a spatial environment is restricted.

Ultimately, while seasonal destination management was utilized as the primary testbed, the cross-platform rendering stability and architectural optimization demonstrated in this study possess significant generalizability. The proposed web-based VR framework serves as a foundational computing model for a variety of critical industrial applications, including remote spatial inspection, real estate documentation, structural digital twin visualization, and Health, Safety, and Environment (HSE) monitoring. Future development should focus on integrating real-time IoT sensor streams into the WebGL environment to further advance the capabilities of interactive spatial systems.

ACKNOWLEDGMENT

This research was financially supported by the Uttaradit Rajabhat University (Fiscal Year 2025).

AUTHORS' CONTRIBUTION

W.M.: conceptualization, methodology, software, investigation, formal analysis, writing—original draft; A.P.: conceptualization, methodology, formal analysis, supervision, writing—original draft, writing—review and editing; P.K.: methodology, investigation, software, visualization, data curation, writing—review and editing.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors used Gemini 3.1 Pro and QuillBot for language refinement and grammatical improvement of the manuscript. These tools were used solely to enhance the clarity and readability of the text. The authors take full responsibility for the integrity of the research content, the accuracy of all data and findings, and the overall scientific contribution of this work.

REFERENCES

[1] X. Wu and I. K. W. Lai, "The use of 360-degree virtual tours to promote mountain walking tourism: Stimulus-organism-response model," Information Technology & Tourism, vol. 24, no. 1, pp. 85-107, 2022.

[2] R. L. Daft and R. H. Lengel, "Organizational information requirements, media richness and structural design," Management Science, vol. 32, no. 5, pp. 554-571, 1986.

[3] U.-K. Lee, "Tourism using virtual reality: Media richness and information system successes," Sustainability, vol. 14, no. 7, p. 3975, 2022.

[4] M. Di Dalmazi, M. Mandolfo, J. Guixeres, M. Alcañiz Raya, and L. Lamberti, "How immersive technologies impact behavioral responses in

destination marketing: The role of physiological arousal, presence, and age," International Journal of Contemporary Hospitality Management, vol. 36, no. 11, pp. 3628-3650, 2024.

[5] R. Yung and C. Khoo-Lattimore, "New realities: A systematic literature review on virtual reality and augmented reality in tourism research," Current Issues in Tourism, vol. 22, no. 17, pp. 2056-2081, 2019.

[6] M. J. Kim, C. K. Lee, and T. Jung, "Exploring consumer behavior in virtual reality tourism using an extended stimulus-organism-response model," Journal of Travel Research, vol. 59, no. 1, pp. 69-89, 2020.

[7] D. Sarkady, L. Neuburger, and R. Egger, "Virtual reality as a travel substitution tool during COVID-19," in Information and Communication Technologies in Tourism 2021, Springer, 2021, pp. 452-463.

[8] I. Bilynets, P. Trkman, and L. Knezevic Cvelbar, "Virtual tourism experiences: Adoption factors, participation and readiness to pay," Current Issues in Tourism, vol. 27, no. 22, pp. 3658-3675, 2024.

[9] M. E. Ali, M. A. Cheema, T. Hashem, A. Ulhaq, and M. A. Babar, "Enabling spatial digital twins: Technologies, challenges, and future research directions," arXiv preprint arXiv:2306.06600, 2023.

[10] X. Tu, J. Autiosalo, R. Ala-Laurinaho, C. Yang, P. Salminen, and K. Tammi, "TwinXR method for using digital twin descriptions in industrial extended reality applications," Frontiers in Virtual Reality, vol. 4, p. 1019080, 2023.

[11] Z. Chen et al., "Splatter-360: Generalizable 360° Gaussian splatting for wide-baseline panoramic images," arXiv preprint arXiv:2412.06250, 2024.

[12] X. Teng, S. Guo, L. Wang, Q. Yang, and F. Deng, "A criterion of radial distortion on structure from motion with consumer-grade cameras," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 48, pp. 275-280, 2022.

[13] T. Kim and F. Biocca, "Telepresence via television: Two dimensions of telepresence may have different connections to memory and persuasion," Journal of Computer-Mediated Communication, vol. 3, no. 2, 1997.

[14] S. Baloglu and K. W. McCleary, "A model of destination image formation," Annals of Tourism Research, vol. 26, no. 4, pp. 868-897, 1999.

[15] C. Flavian, S. Ibanez-Sanchez, and C. Orus, "The impact of virtual, augmented and mixed reality technologies on the customer experience," Journal of Business Research, vol. 100, pp. 547-560, 2019.

[16] I. P. Tussyadiah, D. Wang, T. H. Jung, and M. C. tom Dieck, "Virtual reality, presence, and attitude change: Empirical evidence from tourism," Tourism Management, vol. 66, pp. 140-154, 2018.

[17] S. Brekalo, K. Pap, and B. Trstenjak, "Enhancing rendering performance in complex visualizations by using optimization techniques and algorithms in browser environments," Engineering, Technology & Applied Science Research, vol. 14, no. 3, pp. 14049-14055, 2024.

[18] J. Sun, D. Lenz, H. Yu, and T. Peterka, "Adaptive multi-resolution encoding for interactive large-scale volume visualization through functional approximation," arXiv preprint arXiv:2409.00184, 2024.

[19] O. El-Said and H. Aziz, "Virtual tours a means to an end: An analysis of virtual tours role in changing tourist behavior," Journal of Travel Research, vol. 61, no. 3, pp. 528-548, 2022.

[20] W. Li, X. Chen, and Y. Wang, "How VR technological features prompt tourists' visiting intention: An integrated approach," Sustainability, vol. 15, no. 6, p. 4765, 2023.

[21] N. Rosli, E. R. Johar, H. O. Zaki, and D. F. M. Fernandez, "The rise of virtual tour in tourism: A bibliometric review and future research agenda," International Journal of Academic Research in Business and Social Sciences, vol. 13, no. 2, pp. 781-799, 2023.

[22] V. De Luca, G. Marcantonio, M. C. Barba, and L. T. De Paolis, "A virtual tour for the promotion of tourism of the city of Bari," Information, vol. 13, no. 7, p. 339, 2022.

[23] Y. Zeng, L. Liu, and R. Xu, "The effects of a virtual reality tourism experience on tourist's cultural dissemination behavior," Tourism and Hospitality, vol. 3, no. 1, pp. 314-329, 2022.

[24] Y. Koga and K. Yamamoto, "Tourism support system to utilize virtual reality space reflecting dynamic information in real time," ISPRS International Journal of Geo-Information, vol. 12, no. 8, p. 301, 2023.

[25] L.-L. Tsai, "Factors that influence virtual tourism holistic image: The moderating role of sense of presence," Sustainability, vol. 14, no. 1, p. 467, 2022.

[26] Y. Stappung et al., "Developing 360° virtual tours for promoting tourism in natural parks in Chile," Sustainability, vol. 15, no. 22, p. 16043, 2023.

[27] E. Polishchuk et al., "The theoretical background of virtual reality and its implications for the tourism industry," Sustainability, vol. 15, no. 13, p. 10534, 2023.

APPENDIX A: MEASUREMENT ITEMS AND RELIABILITY

The specific questionnaire items used to operationalize the Human-Computer Interaction (HCI) dimensions, system fidelity, and subsequent behavioral intent are detailed in Table V. The internal consistency for all dimensions exhibited high reliability, with Cronbach's alpha values ranging from 0.905 to 0.985, confirming the robust internal validity of the measurement model.

TABLE V. CONSTRUCTS, MEASUREMENT ITEMS, AND RELIABILITY

Dimension	Indicator	Measurement Items (Scale: 1 = Strongly Disagree to 5 = Strongly Agree)	Cronbach's α
System Fidelity	Virtual Presence	1. I felt as though I was physically present in the virtual environment. 2. The virtual tour created a strong sense of "being there." 3. I felt immersed in the virtual space rather than just looking at a screen.	0.933
	Environ. Realism	1. The destination possesses beautiful natural scenery. 2. The destination has a unique identity distinct from others. 3. I find tourism in this destination to be highly attractive.	0.961
Visual Quality	Visual Fidelity	1. The visual resolution of the 360-degree environment was exceptionally high. 2. The panoramas were crisp, clear, and free from pixelation. 3. The system provided a highly detailed visual experience.	0.985
	Spatial Awareness	1. I gained a clear understanding of the spatial layout of the destination. 2. The system allowed me to accurately perceive the physical scale of the environment.	0.974
Usability	Ease of Navigation	1. Navigating through the virtual environment was intuitive and effortless. 2. I found it easy to move between different spatial nodes.	0.912
	Responsiveness	1. The system responded instantly to my viewing directions and interactions. 2. The virtual tour loaded the high-resolution scenes without noticeable delay.	0.905
Overall	System Effectiveness	1. I intend to search for more information about tourism in this destination. 2. I intend to recommend this destination to others. 3. It is highly likely that I will plan a trip to this destination in the near future.	0.945

Artificial Intelligence in Software System Quality Assurance: A Systematic Literature Review of Techniques, Tools, and Challenges (2016–2026)

Abdullah A H Alzahrani
Department of Computer Engineering and Computing College at Alqunfuda,
Umm Al Qura University, Makkah, Saudi Arabia

***Abstract*—A shift from deterministic testing to artificial intelligence (AI) driven quality ecosystems is necessitated by the rapid evolution of software architectures, and research from 2016 to the present year is combined by this systematic literature review (SLR), so 195 main studies are analyzed, and the path of artificial intelligence in Software Quality Assurance (SQA) is mapped. An enormous course in scholarly output from 2023 onwards is revealed by the findings, and this growth is driven by the industrial adoption of Large Language Models (LLMs) alongside autonomous agentic systems. In addition, three main areas are addressed by this study, and these areas are identified as the taxonomic shift toward multi-agent architectures, the functional effect of AI on labor-intensive activities such as regression testing, and self-healing automation, and the emerging social and technical challenges of ethical governance alongside explainability. Despite incomparable efficiency gains being offered by AI-driven techniques, the industrial success of these tools is strictly limited until the Maintenance Crisis of generated code is resolved and transparency is ensured through Explainable AI (XAI). Finally, the study concludes with a strategic roadmap for Ethical SQA, providing a foundation for future research in autonomous, self-evolving software systems.**

***Keywords*—Software Quality Assurance (SQA); artificial intelligence (AI); systematic literature review (SLR); Large Language Models (LLMs); Explainable AI (XAI)**

I. Introduction

The integration of AI into SQA represents a fundamental shift in how modern software is validated and maintained. As the velocity of AI-assisted development increases, traditional testing methodologies face significant challenges in scaling to meet new requirements for reliability and governance. A broad examination of the current state of SQA is provided within this section of the study, while the primary reasons for the conduction of this systematic review are identified. In addition, particular attention is directed toward the rising difficulties associated with maintenance and the assurance of quality.

A. Overview

The evolution of SQA has reached a critical juncture where traditional deterministic methods are no longer sufficient to secure the complex, distributed architectures of the modern era. As of 2026, the software industry is transitioning from simple automated scripts to AI-driven quality ecosystems capable of autonomous decision-making [1-2]. While previous decades focused on manual inspection and rule-based automation [3], the current landscape is defined by the integration of LLMs and multi-agent systems designed to enhance systemic reliability, safety, and ethical compliance [4-5]. This shift represents a move toward Quality by Design, where AI is not merely an add-on tool but a core component of the software architecting process [6].

B. Motivation

With regards to the motivation for this SLR, the academic investigation is driven by several shifts within the software engineering field. In addition, the software industry is challenged by a Quality Crisis, where human-directed verification capacity has been exceeded by the speed of development assisted by AI [7]. As a result, fundamental digital infrastructure is threatened by the fast shifts, whereby scalability limitations are faced by traditional regression testing, and a transition toward Quality by Design is demanded so that AI is included as a central architectural element instead of a secondary instrument [1, 8].

Furthermore, a Maintenance Crisis has been introduced by the emergence of generative AI. In addition, rigorous validation exceeding the human-written materials is demanded by the substantial volume of code generated by AI, because intelligent vulnerabilities can be introduced at scale [9]. As a result, research is continuously redirected toward autonomously self-healing test automation and agent-based testing [2, 10]. Furthermore, the management of software materials, including bug reports, has been advanced from basic summarization toward advanced neural analysis [11-13], whereby new efforts concerning the ways by which data are combined within the SQA lifecycle are demanded.

However, concerns regarding algorithmic fairness alongside accountability and transparency are introduced by the adoption of these computational models [14]. As a result, a pressing need for a wide combination concerning the current state of XAI within the software development process is continuously maintained so that intelligent quality is retained trustworthy and interpretable [15-19].

C. Research Questions

Regarding the main research goal, this SLR is constructed around three primary research questions (RQs) so that the gap between theoretical AI possibilities and practical SQA realities is resolved. In addition, a separate component concerning the overlap between AI and software quality is designed to be

investigated by each research question, whereby a progression from technical architecture toward applications and ultimately toward the sociotechnical difficulties of adoption is established.

The first research question (RQ1) can be stated as: Which AI and machine learning architectures are currently exploited to enhance software quality, and how can they be categorized within a SQA taxonomy? This question investigates the progression from traditional neural networks toward LLMs and autonomous multiple agent systems [2, 20-21], so that the specific computational frameworks that are most effective for quality tasks are identified.

The second research question (RQ2) can be stated as: What is the measurable impact of AI-driven techniques on fundamental SQA lifecycle activities, and how do these tools address the maintenance overhead of legacy automation? Specifically, this inquiry evaluates empirical evidence regarding efficiency gains in labor-intensive processes [8, 10, 22], such as regression testing and duplicate bug detection, while assessing their ability to maintain or exceed human-level defect detection [9].

The third research question (RQ3) can be stated as: What are the primary technical and ethical barriers hindering the industrial adoption of AI in SQA, and which frameworks exist to ensure transparency and ethical governance? This question combines current research regarding XAI, fairness, and accountability [14-16] to determine how Ethics by Design can be integrated into autonomous testing cycles [5, 23].

D. Research Contributions

A modern multidirectional structure is constructed by this systematic literature review, so the inclusion of artificial intelligence within the software quality assurance field is thoroughly understood, and an exhaustive classification of artificial intelligence architectures is provided by this study so the movement from heuristic models to the advanced application of agentic AI and Small Language Models (SLMs) [2, 20-21] in recent times is accurately mapped. In addition, an exact empirical evaluation of the functional effects of artificial intelligence across the software development life cycle is delivered, and the quality and maintainability of artifacts that are generated by artificial intelligence are contrasted against traditional human-centric methods [9-10] by this analysis. Finally, a strategic plan for ethical software quality assurance is articulated by the review so that XAI methodologies are clearly identified, and the social and technical obstacles to industrial adoption [5, 15-16] are overcome by these methods.

As a result, this structure is offered as an important resource for practitioners so that their testing infrastructure is modernized, and it is also provided to researchers so that the next generation of gaps in autonomous quality engineering are identified.

E. Paper Outline

Regarding the organization of this study, reporting protocols concerning SLR within software engineering are followed so that the structural arrangement is properly governed. In addition, the theoretical foundation is established within Section II through the formal examination of background principles and related literature. Furthermore, the methodology is thoroughly detailed within Section III, whereby the centralized search procedure conducted through the SDL portal alongside the inclusion and exclusion parameters and the required quality evaluation protocol are described. As a result, the gathered outcomes alongside detailed findings are presented within Section IV, and ultimately, the study is concluded within Section V where findings are summarized and future works are proposed.

II. Background and Related Work

In this section, background literature and related academic work are introduced and discussed; therefore, it is required that a clear separation between testing standards and the emerging autonomous models is formally established so that the current developmental path concerning SQA directed by AI is fully understood. In addition, the review in this section is contextualized through the combination of past systematic literature reviews, so that continuous research gaps are clearly clarified. Furthermore, a technical foundation is provided by the detailing of the evolutionary progression concerning AI architectures, crossing from classical machine learning models toward the modern LLMs and multiple agent systems.

A. Theoretical Foundation: AI in Software Quality Assurance

Regarding the theoretical framework of SQA, a major transition from deterministic rule-based automation toward Quality by Design is currently being experienced, whereby AI is incorporated as a central architectural element [1]. Furthermore, this progression is demanded because complex distributed network architectures of the modern era are not able to be secured by traditional methods, and as a result, simple automated scripts are shown to have been surpassed by theoretical principles so that ecosystems directed by AI, which are capable of autonomous decision making, are fully encompassed. In addition, the transition from traditional neural networks and reinforcement learning toward the modern application of LLMs, SLMs, and autonomous multiple agent systems is positioned as central to this theoretical foundation [2, 4], whereby the mathematical and computational frameworks required for advanced tasks, including code summarization, intelligent test composition, and the management of flaky tests through Transformer architectures, are provided by these technologies. However, a flaky test can be defined as a testing scenario by which non-deterministic outcomes are shown when passing and failing results are yielded during multiple executions on the identical software version under an identical environment and configuration [24-27].

Concerning the developmental progression of AI within SQA, a transition from structured knowledge-based systems toward the highly autonomous architectures of the present period is continuously observed. In addition, early technological advancements were heavily maintained by knowledge-based prediction systems, including Bayesian networks, whereby software defect prediction was achieved when fundamental relationships were modeled, and uncertainty was managed through probabilistic reasoning. Furthermore, specialized professional knowledge was allowed to be incorporated within the quality lifecycle by these traditional AI models, whereby a degree of transparency preceding modern explainability requirements was clearly provided to practitioners.

However, although the current technological environment is largely influenced by neural networks and reinforcement learning, continuous relevance is maintained by these operational systems. As a result, Bayesian approaches are continually employed within operational environments where scattered data is observed so that software reliability and risks are assessed. Furthermore, an expanding movement toward hybrid quality ecosystems is shown by the combination of these classical AI methods alongside modern LLMs, whereby both data-driven strength and logical reasoning are applied [28].

B. Related Surveys and Literature Reviews

The existing body of literature includes several systematic reviews that address specific dimensions of AI in software engineering. Early research primarily focused on classical machine learning for defect prediction, while the period between 2019 and 2022 saw a surge in surveys regarding deep learning and ensemble-based modeling. For instance, comprehensive surveys have explored the role of LLMs in the broader software engineering lifecycle and the specific application of chatbots in engineering tasks. Furthermore, the field of XAI has been extensively integrated to address the black-box nature of machine learning models within software development. Other reviews have focused on specialized SQA activities, such as automated test case generation for web and mobile applications and the summarization of software artifacts [4, 11, 15, 18-20, 22].

C. Research Gap

A clear research gap is identified by this study because recent rapid advancements that were achieved after the year 2022 are not collated by any single document, and a large majority of recent scholarly output is generated during this exact period. Furthermore, the collision of the quality crisis and the maintenance crisis is ignored by previous reviews, while individual artificial intelligence tools are merely listed by those same reviews. In addition, the shift toward agentic artificial intelligence and cooperative artificial intelligence systems is inadequately mapped by recent surveys, so a modern collection is required so the gap between theoretical artificial intelligence capabilities and industrial realities is bridged while ethical governance and systems like Defects Prediction as a Service are addressed, specifically regarding ethical governance and frameworks like Defects Prediction as a Service (DePaaS) which is an AI-based Global Unified Software Defect Prediction [7, 9-10, 28].

III. Methodology

A formal systematic literature review protocol is followed by this research, so transparency and reproducibility are ensured for all findings, and the established guidelines that were created by Kitchenham, PRISMA guidelines [36] for software engineering research are strictly applied throughout the entirety of the process. As a result, the methodology is divided into three distinct phases by this structure, and these stages are categorized as the planning phase, where research questions are defined, the execution phase, where the search and data extraction are completed, and the reporting phase, where the evidence is summarized. Furthermore, the initial yield of 3632 records is reduced to a final amount of 195 primary studies by this exact procedure.

A. Research Design

With regards to this research design, this study is organized as a SLR, so that the empirical evidence concerning the effectiveness and application of AI within SQA is built. In addition, the methodology is organized as three phases, including initial preparation where research questions and protocol development are defined, execution where retrieval and selection alongside data extraction are completed, and final reporting where research summarization and analysis are finalized, as it is illustrated within Fig. 1, which was produced by an AI image creation algorithm so that complete academic transparency is protected, whereby the foundational guidelines for software engineering research [29-34] followed.

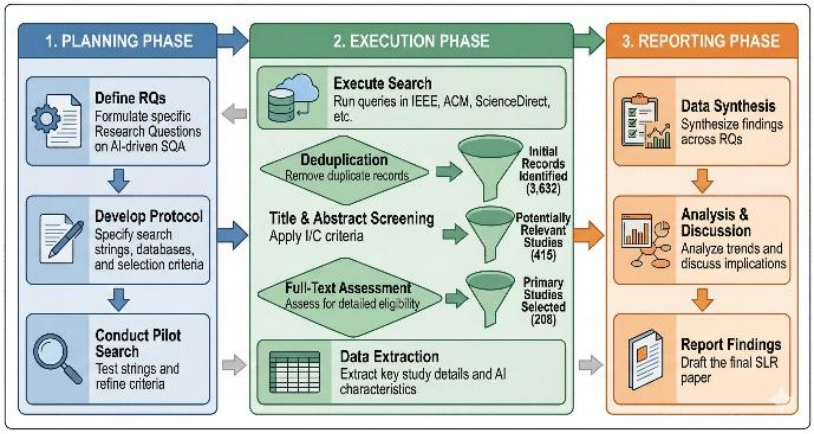


Fig. 1. The SLR workflow.

As a result, the search strategy was executed, so that progressions within the field under consideration over a ten-year period were identified. In addition, primary studies were restricted to the time span between the year 2016 and the present year of 2026, so that it is completely guaranteed that modern AI architectures, including Transformers and multiple agent systems are reflected by the review. However, it must be recognized that important earlier publications preceding this designated timeframe are also cited by the manuscript within the background and methodology sections. As a result, these references are carefully included so that the historical and theoretical background of SQA is properly established.

B. Search Strategy and Data Sources

With regards to the centralized search process, the Saudi Digital Library (SDL) is employed as the main research gateway and interface so that an extensive search is guaranteed.

In addition, the definitive temporal boundary concerning the captured literature is established by the final execution of the refined search string, which was completed on 19 January of the present year 2026. Furthermore, peer-reviewed publications released between the year 2016 and early 2026, alongside high-impact preprints, are encompassed by this search. Additionally, combined access concerning several high-impact digital libraries is provided by the SDL portal, whereby a simultaneous query is allowed across multiple academic databases, including IEEE Xplore and the ACM Digital Library, alongside ScienceDirect and SpringerLink. Moreover, a discovery layer is provided by the SDL, by which referential databases like Scopus and Web of Science are covered. Furthermore, it is guaranteed by this broad indexing ability of SDL that metadata and citations from a varied selection of software engineering and AI literature are captured, whereby the reach of the review is extended beyond individual publisher repositories.

Regarding the search strategy, field-related abbreviations supported by the SDL interface are employed so that results are refined and high precision is ensured. In addition, the TI (Title) and AB (Abstract) tags are applied by the procedure. As a result, the search is constrained by the use of these qualifiers toward papers where SQA and AI are positioned at the center of the research. However, papers where these topics are mentioned incidentally within the full text are officially excluded from the results. In addition, a deliberate methodological decision was finalized so that broad search engines like Google Scholar generating an excessive volume of irrelevant results, are avoided, and instead, a controlled database environment, SDL interface, is strictly prioritized for the collection phase.

With regards to the search string, Boolean operators are used so that terms connected to the subjects SQA, AI, and SLR are combined, whereby strict alignment alongside established systematic mapping techniques within the field is guaranteed [35]. In addition, although a conceptual overview concerning search logic is provided in Fig. 2, which is also generated by an AI image creation algorithm so that complete academic transparency is maintained, the exact syntax and Boolean operators required by different digital libraries are supplied within Table I. As a result, when both the visual logic and the technical execution are presented, it is ensured that the methodology is made accessible and is reproducible.

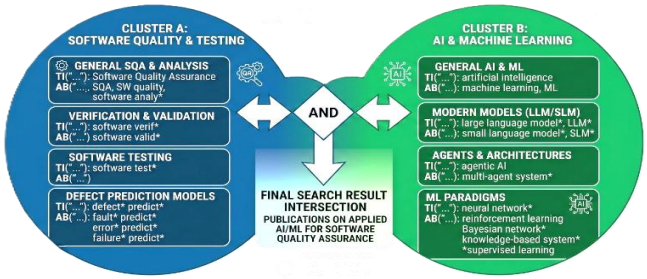


Fig. 2. Schematic representation of the SDL search string.

TABLE I. SDL SEARCH STRING

Search String	((TI("software quality assurance") OR TI("SQA") OR TI("SW quality") OR TI("software test*") OR TI("software verif*") OR TI("software valid*") OR TI("software analy*") OR TI("defect* predict*") OR TI("fault* predict*") OR TI("error* predict*") OR TI("failure* predict*")) OR (AB("software quality assurance") OR AB("SQA") OR AB("SW quality") OR AB("software test*") OR AB("software verif*") OR AB("software valid*") OR AB("software analy*") OR AB("defect* predict*") OR AB("fault* predict*") OR AB("error* predict*") OR AB("failure* predict*"))) AND ((TI("AI") OR TI("artificial intelligence") OR TI("machine learning") OR TI("ML") OR TI("large language model*") OR TI("LLM*") OR TI("small language model*") OR TI("SLM*") OR TI("agentic AI") OR TI("multi-agent system*") OR TI("neural network*") OR TI("reinforcement learning") OR TI("Bayesian network*") OR TI("knowledge-based system*") OR TI("*supervised learning")) OR (AB("AI") OR AB("artificial intelligence") OR AB("machine learning") OR AB("ML") OR AB("large language model*") OR AB("LLM*") OR AB("small language model*") OR AB("SLM*") OR AB("agentic AI") OR AB("multi-agent system*") OR AB("neural network*") OR AB("reinforcement learning") OR AB("Bayesian network*") OR AB("knowledge-based system*") OR AB("*supervised learning")))

C. Inclusion and Exclusion Criteria

To filter the initial search results into a set of primary studies that directly address the research questions, the inclusion and exclusion criteria were applied, as detailed in Table II.

TABLE II. INCLUSION AND EXCLUSION CRITERIA FOR STUDY SELECTION

Criterion	Inclusion (IC)	Exclusion (EC)
Timeframe	Published between 2016 and January 2026.	Published prior to 2016.
Peer Review	Journal articles, conference proceedings, and high-impact pre-prints (2025–2026).	Grey literature, blog posts, and white papers.
Content Focus	Primary focus on AI/ML applied to SQA tasks.	Focus on hardware QA, robotics, or general AI theory.
Language	Full-text available in English.	Non-English publications.
Quality	Clearly defined methodology and results.	Short papers (< 4 pages) or abstract-only entries.

D. Study Selection Process

Concerning the selection process, the PRISMA 2020 protocol was strictly followed so that transparency and replicability were ensured [36], and in addition, automated system filtering within the Saudi Digital Library (SDL) was employed by the procedure so that preliminary duplication removal was executed and metadata completeness was formally validated. Furthermore, a manual three-stage review was formally conducted, whereby titles and abstracts were initially screened during the first stage so that clearly irrelevant entries were eliminated, and as a result, a full-text review was performed during the second stage so that alignment alongside all inclusion criteria was completely guaranteed.

Furthermore, a structured set of strict rules and measurable parameters was established so that the academic validity of each retrieved study was objectively judged and potential bias was thoroughly assessed during the final screening procedure. In addition, a standardized scoring system was adopted by which

each manuscript was evaluated concerning research design clarity and empirical validity, so that arbitrary decisions were avoided and only studies where high methodological rigor was observed were retained for the literature review. As a result, any publications where insufficient empirical evidence was detected were officially excluded so that the internal validity of the research was protected.

E. Data Extraction and Composition

With regards to the extraction process, a uniform data extraction form was employed so that key information from the selected primary studies was successfully captured, whereby the exact AI architecture applied, the targeted SQA activity, and the reported performance measurements were reported. Furthermore, the extracted data were sequentially combined through narrative composition and objective analysis so that the research questions defined within Section I C were thoroughly answered, whereby strict consistency alongside methodologies adopted within previous large-scale software engineering surveys was maintained [4].

IV. Results and Discussion

In this section, a multidimensional assembly of the 195 primary studies found through the systematic search process is introduced. In addition, through the assessment of data against the main research questions, structural directions, operational influences on the SQA lifecycle, and the social-technical governance needed for dependable AI are thoroughly examined. Furthermore, the succeeding subsections supply a longitudinal review showing how AI methodologies and tools have advanced to satisfy industrial needs regarding safety and reliability between the year 2016 and the present year.

A. Bibliometric Distribution and Temporal Trends

With regard to the study selection procedure, two distinct phases were completed so that an exhaustive collection of literature was achieved. In addition, 1416 results were produced by using the search string, where central AI and SQA terms were targeted. However, so that possible confirmation bias was evaded, the search strategy was expanded so that wildcards and plural forms alongside a wider selection of classical AI architectures like Bayesian networks and reinforcement learning were included. Furthermore, a total of 3632 results were yielded by this refined execution of the refined search string, shown in Table I, within the Saudi Digital Library interface, as it is detailed within Appendix A (see Fig. 4). As a result, it is guaranteed by the inclusion of these wider expressions that the early basic stages of AI within SQA are proven exact and are not concealed by modern generative AI terminology. In addition, so that the strict accuracy of data extraction was completely protected, the results were exported and verified through both RIS and BibTeX formats, whereby it was guaranteed that zero records were lost because of interface-specific exportation errors.

Furthermore, upon exportation and internal validation, the gathered papers collection was narrowed to 2128 records. In addition, a secondary manual duplication removal protocol was applied so that overlapping indexing across databases like IEEE Xplore and the ACM Digital Library were addressed, whereby the deletion of 1050 redundant entries were achieved. However, a substantial record connected to non-software areas like medical diagnostics and hardware-centered quality assurance was identified by the screening phase. As a result, after irrelevant studies were excluded, the final main study set was confirmed at 195 unique publications, as it is detailed within Table VII of Appendix B. In addition, the “Show More Results”, as can be seen in Appendix A (Fig. 4), pagination displaying additional results was fully exhausted during the identification phase so that it was guaranteed that all indexed items were captured. Ultimately, it is confirmed by this distribution which is detailed within Table III that although terminology bias was eliminated by the expanded search, the central research output directed toward software is maintained as consistent.

TABLE III. Study Selection and Filtering Results

Phase	Action	Records Remaining
Identification	Initial Search Hits (Expanded Query)	3,632
System Filtering	Records Exported (Validation & Internal Deduplication)	2,128
Manual Deduplication	Removal of overlapping records across digital libraries	1,078
Domain Screening	Exclusion of non-software QA (Medical, Hardware, etc.)	432
Final Inclusion	**Refined primary study**	**195**

The temporal analysis of the final primary study set (N=195) indicates an evolving research trajectory within the domain of artificial intelligence in SQA. The period from 2016 to 2018 represents a foundational stage, contributing only 2% of the total corpus and focusing primarily on classical machine learning for defect prediction. A transitional growth phase is observed between 2019 and 2022, characterized by the rising integration of deep learning architectures and ensemble-based modeling. The most significant surge in scholarly output occurred from 2023 onwards, with over 80% of the primary studies published within the final three years of the inclusion timeframe. This exponential growth reflects a paradigm shift driven by the industrial adoption of LLMs, Transformers, and autonomous agentic systems. Notably, the data for 2026 represents a preliminary and incubatory phase of research output, reflecting studies indexed within the first month of the year. While these 2026 results represent only a partial annual cycle, they provide critical early insights into the emergence of agentic AI and self-healing frameworks. The annual distribution and corresponding research milestones are presented in Table IV and Fig. 3.

TABLE IV. Annual Distribution and Research Milestones in SQA-AI (2016–2026) for the Primary Study Set.

Year Interval	Study Count (n)	Percentage (%)	Key Research Milestone
2016–2018	3	2%	Foundations of Machine Learning for Defect Prediction
2019–2020	8	4%	Early Adoption of Deep Learning and CNNs
2021–2022	20	10%	Maturity of Ensemble Methods and Test Optimization
2023–2024	97	50%	Shift to LLMs, Transformers, and Generative SQA

2025–2026*1	67	34%	Emergence of Agentic AI and Explainable Governance
Total	**195**	**100%**	—

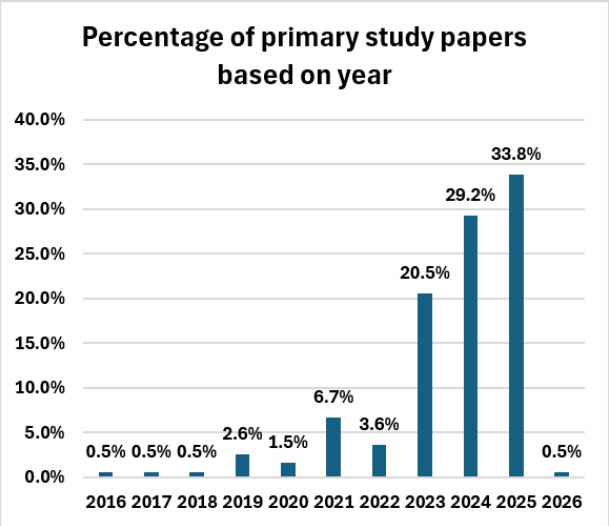


Fig. 3. Percentage of primary study papers based on the year.

With regard to the gathered data shown earlier, it is confirmed that the most active developmental phase is currently being experienced by the academy. In addition, a movement from theoretical AI capabilities toward applied industrial solutions is clearly indicated by the volume of studies published during recent years (2023–2026), whereby heightened attention is directed toward resolving the Maintenance Crisis within older automation systems alongside the requirement that XAI is employed so that ethical governance is entirely guaranteed within mission-critical software environments.

However, although peer-reviewed literature is strictly prioritized by the academic review, the inclusion of selected preprints from repositories, including arXiv, was dictated by the rapid emergence of agentic AI during late 2025. As a result, essential evidence concerning the current state-of-the-art is securely provided by these academic sources, whereby they were formally included only when an equivalent peer-reviewed version was not available.

1) Longitudinal evolution of research inquiries: With regard to the major shifts within SQA, it is required that academic milestones are viewed as an overlapping range rather than isolated intervals so that these structural changes are properly defined. Furthermore, this progression is illustrated by Table V through the mapping of the annual distribution and academic milestones across the primary study set, whereby the movement from basic defect prediction toward autonomous systems is clearly emphasized. As a result, the exact manner by which escalating structural complications have been addressed by the industry is revealed by the progression of each individual research question:

- RQ1 (Architectures): It is observed that research was transitioned from classical machine learning concerning defect prediction between the years 2016 and 2018 toward deep learning and combined methodologies between the years 2019 and 2022. As a result, central attention was directed toward LLMs by the year 2023, and ultimately, multiple agent systems are currently being deployed so that autonomous composition is fully achieved.
- RQ2 (Functional Impact): Early research efforts were directed toward the improvement of labor-intensive tasks, including regression testing. However, because the Quality Crisis was initiated by the extreme speed of development assisted by AI, academic research was formally re-oriented toward autonomously self-healing test automation between the year 2024 and the current period so that the resulting Maintenance Crisis concerning code generated by AI is managed.
- RQ3 (Socio-technical Governance): Although AI was considered a black-box system by early academic studies, a massive increase in XAI and ethical design principles was experienced during the period following the year 2023. As a result, continuous academic efforts are represented by this progression so that AI decisions are aligned with human testing assumptions.

TABLE V. OVERLAPPING RESEARCH MILESTONES AND CRISIS RESPONSES.

Period	Primary Focus (RQ1/RQ2)	Crisis Context	Strategic Response (RQ3)
2016–2020	Classical ML & Early CNNs	Foundational Scalability	Rule-based Transparency
2021–2024	Transformers & Generative SQA	Quality Crisis	XAI & Developer Trust
2025–2026	Agentic AI & Self-Healing	Maintenance Crisis	Standardized Governance (DePaaS)

B. RQ1: Taxonomic Mapping of AI Architectures in SQA

With regards to the architectural environment of AI within SQA, a major transition from usual machine learning toward advanced deep learning and generative frameworks has been experienced [37-39], and furthermore, the deployment of LLMs and Transformer architectures is centralized within current research, whereby high level cognitive automation including the generation of unit tests and the progression of test oracle automation is executed by these technologies [25-27]. As a result, transformer architectures are proposed as strong solutions concerning Graphical User Interface (GUI) testing so that the difficulties associated with automated test case generation and the management of flaky tests are resolved [24, 40-41].

In addition, researchers employ sophisticated hybrid techniques alongside generative models so that predictive precision is increased, and as a result, this practice is illustrated by the creation of new techniques, including VoStackSDD concerning software defect density prediction [42] alongside the application of bootstrap aggregation Bagging so that internal

[1]*Note: 2026 data reflects results collected up to January 19, 2026, and represents an incubatory stage of the current research cycle.

project defect prediction is improved [43]. Furthermore, deep learning architectures are maintained as essential components concerning functional automation, including web user interface testing based upon Selenium UI testing [40-41] alongside specialized multi-level signal quality classification, while the role of reinforcement learning concerning software testing composition alongside the examination of quantum reinforcement learning is included by newly visible classification borders. [44-46].

C. RQ2: Impact on Core SQA Lifecycle Activities

With regards to the software development life cycle, labor-intensive and error-prone tasks are greatly improved by methods and tools directed by AI, and in addition, machine learning models are adopted by researchers within the area of defect prediction and localization, so that source code syntax patterns are used to examine bug-inducing cases. Furthermore, these efforts are evaluated across internal and external project environments so that scalability and general effectiveness are accurately determined [43, 47-49].

As a result, the SQA lifecycle is additionally affected by continuous improvements concerning test case management and maintenance, whereby improvement initiatives are extended toward test suite reduction. Furthermore, unsupervised machine learning approaches are employed by researchers within this area so that redundant testing scenarios are eliminated while extensive software coverage is maintained [50]. Additionally, modern AI tools are being integrated into debugging and log analysis processes [38], while LLMs are being paired with stochastic science to enhance path coverage and crash localization during fuzz testing [51].

It can be concluded that labor-intensive and error-prone tasks across the software development life cycle are refined by artificial intelligence-driven techniques, and aggregated performance statistics have been compiled so that empirical claims are supported. In addition, within the field of defect prediction and localization, machine learning models are applied so that source code syntax patterns are analyzed, and bug-inducing commits are identified [47], while accuracy ranges between 80% and 95% are reported across the evaluated literature. Furthermore, these methodologies are evaluated across within-project and cross-project contexts so that scalability is determined [43, 48-49], and a median F-measure distribution of 87% is observed by the surveyed studies. The software testing lifecycle is affected by advancements in test case management, and unsupervised machine learning approaches are employed so that redundant test cases are eliminated [50] and test execution time is decreased by an average of 40%. Finally, modern artificial intelligence tools are applied to debugging processes [38], and large language models are combined with stochastic science so that path coverage is expanded and crash localization is improved during fuzz testing [51], which is detailed in Table VI.

TABLE VI. AGGREGATED PERFORMANCE METRICS FOR AI-DRIVEN SQA LIFECYCLE ACTIVITIES.

Performance Metric	Reported Statistical Range	Targeted Software Lifecycle Activity
Defect Prediction Accuracy	between 80% and 95%	Bug-inducing commit identification
F measure Distribution	median F measure distribution of 87%	Within project and cross project localization
Test Execution Time	decreased by an average of 40%	Redundant test suite elimination

D. RQ3: Trustworthiness, Explainability, and Ethical Governance

To facilitate full-scale industrial adoption, recent research addresses the critical barriers of AI's black-box nature through both interpretability and explainability. While explainability is frequently employed to clarify the causal relationships between inputs and outputs (often through post-hoc techniques), interpretability remains a more stringent requirement for mission-critical SQA. It is worth noting here that XAI can be defined as a precise set of processes and methods by which the exact reasoning behind algorithmic decisions is made fully transparent to human operators [17, 52]. Interpretability allows developers to identify and understand the specific contributions of a model's internal components to its decision-making process. Specifically, in safety-critical domains such as aerospace, providing a human-readable explanation is often insufficient; instead, researchers are pivoting toward intrinsically interpretable architectures, such as neural additive models [49], which provide a transparent view of how individual features influence a defect prediction. [52]. This transparency is particularly vital for mission-critical systems (e.g., safety-critical aerospace or medical software), where a lack of software quality can lead to catastrophic events, including loss of life or significant environmental damage. A primary technical hurdle in mission-critical SQA is the management of false positives; industrial studies explore how developers perceive AI-generated defects [53] and how chatbot decisions compare to human testers' assumptions [54-55].

Furthermore, as industries are constantly oriented toward development assisted by AI, considerable attention regarding the quality assurance of code generated by AI is identified by this review, so that a Maintenance Crisis is avoided. In addition, this issue is resolved by recent frameworks through the implementation of autonomously self-healing automation, whereby autonomous agents are employed so that damaged testing scripts are detected and repaired immediately, while it is ensured that rigorous reliability examinations are not bypassed by fast deployment cycles. With regards to administrative protocols, governance actions are currently being formalized through frameworks including DePaaS, recognized as a global unified software defect prediction directed by AI, whereby central attention is given to feasibility analysis and standardized routines concerning worldwide software environments [28]. As a result, it is emphasized by state-of-the-art studies that the trustworthiness of autonomous agents seriously depends on data quality and complexity, whereby informative defect prediction methods are required [56-58].

E. Research Summarization and Industrial Implications

With regard to literature reviews, unparalleled automation capacities concerning unit testing and oracles are suggested to be provided by AI architectures, including LLMs and Transformers. In addition, it is argued that success concerning these technologies is linked to the resolution of the Maintenance Crisis alongside overcoming the difficulty which is created by the verification and preservation of massive volumes of code

which is generated by AI. Furthermore, the creation of self-healing frameworks is encouraged by recent studies so that testing scripts are repaired independently and fast development cycles directed by AI are matched. However, a movement toward XAI is considered mandatory for operational systems within healthcare or space environments where severe events, including human casualties or major service interruptions, are caused by poor software quality. As a result, transparency regarding defect prediction is demanded within these risk environments so that predictable safety boundaries are maintained.

Furthermore, the attention of industrial practitioners must be redirected from the accuracy of computational models toward the wise management of developer trust. In addition, this exact objective is achieved through the reduction of false positive results alongside the execution of formal verification, while strict alignment between decisions made by AI and with human assumptions is ensured [59]. As a result, aggressive commercial delivery constraints are successfully satisfied by organizations through the adoption of these rigorous validation methods, while the strict integrity of infrastructure is entirely protected. At last, the inclusion of Ethics by Design principles through standardized governance frameworks such as DePaaS provides guidelines for ensuring that autonomous SQA agents operate within safe and predictable boundaries.

Experimental testing phases are recognized to have been exceeded by current state-of-the-art studies, whereby the incorporation of AI within SQA is now seen as an absolute need, although a severe gap is constantly observed regarding the automated validation of materials that are generated by AI. Since this gap is not yet resolved, a transition toward agent-directed SQA is identified, whereby multiple agent systems are deployed so that the cross-validation of programming and testing scenarios is thoroughly executed. In addition, when SLMs are combined for specialized verification alongside XAI structures, the Quality Crisis can be cured through transparent feedback cycles, whereby SQA is changed from a reactive obstruction toward a proactive self-healing infrastructure so that industrial demands regarding both developmental speed and reliability are completely satisfied.

V. Conclusion and Future Work

A. Conclusion

In conclusion, a total of 195 primary studies have been combined within this systematic literature review so that the changing adoption of AI within SQA during a recent 10-year period can be mapped, and it is shown by the results that the most active period is currently being experienced by this field, given that greater than 80% of primary studies have been produced since the year 2023. In addition, this increase is recognized as a clear exit from conventional rule-based automation toward self-governing quality environments that are sustained by LLMs, Transformers, and multiple agent systems. Although considerable abilities in the automation of unit test generation, oracle automation, and defect localization are supplied by these AI-driven architectures, industrial deployment is restricted by a combined Quality Crisis and Maintenance Crisis. In particular, human verification has been surpassed by the extreme speed of computer-assisted development, meaning that a transition toward agent-based SQA and the installation of SLMs for localized, efficient verification is required, whereby it is guaranteed by these Quality by Design methods that AI is maintained as a central, controllable structural element. Finally, it is concluded by this investigation that a transition toward XAI and standardized governance structures, including DePaaS, is absolutely mandatory so that the developer reliance required within operational environments can be established.

B. Future Work

As the distance between theoretical AI capabilities and practical industrial realities is reduced, the simple classification of tools must be overcome by future investigations so that the maintenance crisis is fully resolved. For immediate objectives, priority must be assigned to the development of autonomously correcting structures or self-healing frameworks that are designed so that the severe administrative burdens caused by AI-generated code are effectively lessened. Concurrently, the socio-technical aspects of programmer trust must be examined through the continuous advancement of XAI methodologies. In particular, the reduction of false positive results must be targeted by these procedures so that exact alignment between AI-driven decisions and human expectations is completely guaranteed. Additionally, a necessity that quantum reinforcement learning and agentic collaboration are closely investigated is proposed by newly visible classification borders in order to manage the complexity of distributed architectures. In conclusion, a need is continually maintained by the formalization of ethical design principles within SQA, where standardized governance structures like DePaaS are applied so that safe and legally compliant operational boundaries for autonomous agents are enforced.

Acknowledgment

Sincere gratitude is formally expressed to the Umm Al Qura University because an invaluable opportunity was provided so that academic research regarding the software engineering field could be conducted, while heartfelt thanks are similarly extended to family members and friends.

References

[1] M. Islam, F. Khan, S. Alam, and M. Hasan, “Artificial Intelligence in Software Testing: A Systematic Review,” in TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), Oct. 2023, pp. 524–529. doi: 10.1109/TENCON58879.2023.10322349.

[2] S. Naqvi, M. Baqar, and N. A. Mohammad, “The Rise of Agentic Testing: Multi-Agent Systems for Robust Software Quality Assurance,” Jan. 05, 2026, arXiv: arXiv:2601.02454. doi: 10.48550/arXiv.2601.02454.

[3] M. Last, A. Kandel, and H. Bunke, Artificial Intelligence Methods in Software Testing, vol. 56. in Series in Machine Perception and Artificial Intelligence, vol. 56. WORLD SCIENTIFIC, 2004. doi: 10.1142/5549.

[4] X. Hou et al., “Large Language Models for Software Engineering: A Systematic Literature Review,” ACM Trans. Softw. Eng. Methodol., vol. 33, no. 8, pp. 1–79, Nov. 2024, doi: 10.1145/3695988.

[5] M. Ramachandran, “Quality Assurance Framework for AI Ethics by Design,” in Engineering Ethics of AI by Design, Singapore: Springer Nature Singapore, 2026, pp. 623–687. doi: 10.1007/978-981-95-2909-4_12.

[6] A. Ahmad, M. Waseem, P. Liang, M. Fahmideh, M. S. Aktar, and T. Mikkonen, “Towards Human-Bot Collaborative Software Architecting with ChatGPT,” in Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, Oulu Finland: ACM, Jun. 2023, pp. 279–285. doi: 10.1145/3593434.3593468.

[7] R. Karanjkar and S. Phalke, "AI Investment and the Pendulum Effect: The Crisis of Software Quality & Infrastructure," Int. J. Res. Appl. Innov., vol. 8, no. 6, 2025, doi: 10.15662/IJRAI.2025.0806018.

[8] M. M. Alam, S. I. Priti, S. Ahmed, K. Fatema, M. Hasan, and N. Nahar, "AI-Driven Solutions for Regression Testing: Insights from Bangladesh Software Industry," in Information Systems for Intelligent Systems, C. S. In, N. S. Londhe, N. Bhatt, and M. Kitsing, Eds., Singapore: Springer Nature, 2025, pp. 495–508. doi: 10.1007/978-981-96-1210-9_43.

[9] T. M. Eltabakh, N. Nabil Soudi, and D. Shawky, "Quality of AI-Generated vs. Human-Generated Code," in 2024 34th International Conference on Computer Theory and Applications (ICCTA), Dec. 2024, pp. 200–205. doi: 10.1109/ICCTA64612.2024.10974782.

[10] S. C. P. Saarathy, S. Bathrachalam, and B. K. Rajendran, "Self-Healing Test Automation Framework using AI and ML," Int. J. Strateg. Manag., vol. 3, no. 3, pp. 45–77, 2024, https://doi.org/10.47604/ijsm.2843

[11] N. Nazar, Y. Hu, and H. Jiang, "Summarizing software artifacts: A literature review," J. Comput. Sci. Technol., vol. 31, no. 5, pp. 883–909, 2016, https://doi.org/10.1007/s11390-016-1671-1

[12] G. Long, "Enhancing automated bug report analysis through advanced neural language models," 2024, https://doi.org/10.26174/thesis.lboro.27859332.

[13] G. Long, J. Gong, H. Fang, and T. Chen, "Learning Software Bug Reports: A Systematic Literature Review," ACM Trans. Softw. Eng. Methodol., p. 3750040, Jul. 2025, doi: 10.1145/3750040.

[14] V. Bogina, A. Hartman, T. Kuflik, and A. Shulner-Tal, "Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics," Int. J. Artif. Intell. Educ., vol. 32, no. 3, pp. 808–833, Sep. 2022, doi: 10.1007/s40593-021-00248-0.

[15] S. Cao et al., "A Systematic Literature Review on Explainability for ML/DL-based Software Engineering," ACM Comput. Surv., vol. 58, no. 4, pp. 1–34, Mar. 2026, doi: 10.1145/3763230.

[16] S. Cao et al., "A Systematic Literature Review on Explainability for Machine/Deep Learning-based Software Engineering Research," Feb. 05, 2025, arXiv: arXiv:2401.14617. doi: 10.48550/arXiv.2401.14617.

[17] T. Clement, N. Kemmerzell, M. Abdelaal, and M. Amberg, "a systematic metareview of explainable AI (XAI) aligned to the software development process," Mach. Learn. Knowl. Extr., vol. 5, no. 1, pp. 78–108, 2023, https://doi.org/10.3390/make5010006

[18] A. H. Mohammadkhani, N. S. Bommi, M. Daboussi, O. Sabnis, C. Tantithamthavorn, and H. Hemmati, "A Systematic Literature Review of Explainable AI for Software Engineering," Feb. 13, 2023, arXiv: arXiv:2302.06065. doi: 10.48550/arXiv.2302.06065.

[19] A. Haji Mohammadkhani, "Explainable AI for Software Engineering: A Systematic Review and an Empirical Study," 2023, Accessed: Jan. 19, 2026, https://dx.doi.org/10.11575/PRISM/40697

[20] F. Wang et al., "A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness," ACM Trans. Intell. Syst. Technol., vol. 16, no. 6, pp. 1–87, Dec. 2025, doi: 10.1145/3768165.

[21] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, and Q. Wang, "Software testing with large language models: Survey, landscape, and vision," IEEE Trans. Softw. Eng., vol. 50, no. 4, pp. doi: 911–936, 2024, 10.1109/TSE.2024.3368208

[22] S. Gupta and S. K. Gupta, "A systematic study of duplicate bug report detection," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 1, 2021, DOI: 10.14569/IJACSA.2021.0120167

[23] A. Patil, "Advancing Software Quality: A Standards-Focused Review of LLM-Based Assurance Techniques," ArXiv Prepr. ArXiv250513766, 2025, https://doi.org/10.48550/arXiv.2505.13766

[24] Z. Khaliq, S. U. Farooq, and D. A. Khan, "A deep learning-based automated framework for functional User Interface testing," Inf. Softw. Technol., vol. 150, Oct. 2022, doi: 10.1016/j.infsof.2022.106969.

[25] D. M. Zapkus and A. Slotkienė, "Unit Test Generation Using Large Language Models: A Systematic Literature Review.," Vilnius Univ. Open Ser., pp. 136–144, Mar. 2024, doi: 10.15388/LMITT.2024.20.

[26] D. M. Zapkus and A. Slotkienė, "Quality Evaluation of Large Language Models Generated Unit Tests: Influence of Structured Output.," Vilnius Univ. Open Ser., pp. 281–288, Jun. 2025, doi: 10.15388/mitt.2025.32.

[27] F. Molina, A. Gorla, and M. d'Amorim, "Test Oracle Automation in the Era of LLMs.," ACM Trans. Softw. Eng. Methodol., vol. 34, no. 5, pp. 1–24, Jun. 2025, doi: 10.1145/3715107.

[28] Mahesha Pandit et al., "Towards Design and Feasibility Analysis of DePaaS: AI Based Global Unified Software Defect Prediction Framework," Appl. Sci., vol. 12, no. 1, pp. 493–493, Jan. 2022, doi: 10.3390/app12010493.

[29] B. Kitchenham, "Procedure for undertaking systematic reviews," Comput. Sci. Depart-Ment Keele Univ. TRISE-0401 Natl. ICT Aust. Ltd 0400011T 1 Jt. Tech. Rep., 2004. Keele University Technical Report TR/SE-0401 ISSN:1353-7776

[30] B. Kitchenham, "Procedures for performing systematic reviews," Keele UK Keele Univ., vol. 33, no. 2004, pp. 1–26, 2004.

[31] B. A. Kitchenham, D. Budgen, and P. Brereton, Evidence-based software engineering and systematic reviews, vol. 4. CRC press, 2015. ISBN:978-1-4822-2865-6

[32] B. A. Kitchenham, T. Dyba, and M. Jorgensen, "Evidence-based software engineering," in Proceedings. 26th International Conference on Software Engineering, IEEE, 2004, pp. 273–281. DOI: 10.1109/ICSE.2004.1317449

[33] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," Inf. Softw. Technol., vol. 51, no. 1, pp. 7–15, 2009, https://doi.org/10.1016/j.infsof.2008.09.009

[34] B. Kitchenham et al., "Systematic literature reviews in software engineering–a tertiary study," Inf. Softw. Technol., vol. 52, no. 8, pp. 792–805, 2010, https://doi.org/10.1016/j.infsof.2010.03.006

[35] H. Sofian, N. A. M. Yunus, and R. Ahmad, "Systematic mapping: Artificial intelligence techniques in software engineering," IEEE Access, vol. 10, pp. 51021–51040, 2022, DOI: 10.1109/ACCESS.2022.3174115

[36] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and The PRISMA Group, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," PLoS Med., vol. 6, no. 7, p. e1000097, Jul. 2009, doi: 10.1371/journal.pmed.1000097.

[37] Peter Kokol, "The Use of AI in Software Engineering: A Synthetic Knowledge Synthesis of the Recent Research Literature," Information, vol. 15, no. 6, pp. 354–354, Jun. 2024, doi: 10.3390/info15060354.

[38] M. Staron, S. Abrahao, G. Gay, and A. Serebrenik, "Testing, Debugging, and Log Analysis With Modern AI Tools," IEEE Softw., vol. 41, pp. 99–102, Jan. 2024, doi: 10.1109/MS.2023.3339408.

[39] M. Staron and S. Abrahão, "Hybrid Classical-AI Systems for Software Testing and Bug Fixing," IEEE Softw., vol. 42, no. 6, pp. 106–110, Jan. 2025, doi: 10.1109/MS.2025.3597698.

[40] [Z. Khaliq, D. A. Khan, and S. U. Farooq, "Using deep learning for selenium web UI functional tests: A case-study with e-commerce applications," Eng. Appl. Artif. Intell., vol. 117, Jan. 2023, doi: 10.1016/j.engappai.2022.105446.

[41] Z. Khaliq, S. U. Farooq, and D. A. Khan, "Transformers for GUI Testing: A Plausible Solution to Automated Test Case Generation and Flaky Tests.," Comput. 00189162, vol. 55, no. 3, pp. 64–73, Mar. 2022, doi: 10.1109/MC.2021.3136791.

[42] J. Kaur, A. Kaur, and K. Kaur, "VoStackSDD: A Novel Ensemble Technique for Software Defect Density Prediction.," J. Commun. Softw. Syst., vol. 21, no. 3, pp. 306–316, Sep. 2025, doi: 10.24138/jcomss-2025-0006.

[43] U. S. Bhutamapuram and R. Sadam, "With-in-project defect prediction using bootstrap aggregation based diverse ensemble learning technique," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 10, pp. 8675–8691, Nov. 2022, doi: 10.1016/j.jksuci.2021.09.010.

[44] Francisco David Pérez Reynoso, Jorge Alberto Soto Cajiga, Luis Alberto Gordillo Roblero, and Paola Andrea Niño Suárez, "Toward Embedded Multi-Level Classification of 12-Lead ECG Signal Quality Using Spectrograms and CNNs," Appl. Sci., vol. 15, no. 24, pp. 12976–12976, Dec. 2025, doi: 10.3390/app152412976.

[45] A. Abo-eleneen, A. Palliyali, and C. Catal, "The role of Reinforcement Learning in software testing," Inf. Softw. Technol., vol. 164, Dec. 2023, doi: 10.1016/j.infsof.2023.107325.

[46] S. Park and J. Kim, "Trends in quantum reinforcement learning: State-of-the-arts and the road ahead.," ETRI J., vol. 46, no. 5, pp. 748–758, Oct. 2024, doi: 10.4218/etrij.2024-0153.

[47] M. Nadim and B. Roy, "Utilizing source code syntax patterns to detect bug inducing commits using machine learning models.," Softw. Qual. J., vol. 31, no. 3, pp. 775–807, Sep. 2023, doi: 10.1007/s11219-022-09611-3.

[48] Y. Z. Bala, P. A. Samat, K. Y. Sharif, and N. Manshor, "The influence of machine learning on the predictive performance of cross-project defect prediction: empirical analysis.," Telkomnika, vol. 22, no. 4, pp. 830–837, Aug. 2024, doi: 10.12928/TELKOMNIKA.v22i4.25916.

[49] Ruiqi He, Yong Li, and Chi Sun, "Understanding Software Defect Prediction Through eXplainable Neural Additive Models," IEEE Access, vol. 13, pp. 82860–82873, Jan. 2025, doi: 10.1109/ACCESS.2025.3567140.

[50] A. Sebastian, H. Naseem, and C. Catal, "Unsupervised Machine Learning Approaches for Test Suite Reduction.," Appl. Artif. Intell., vol. 38, no. 1, pp. 1–31, Dec. 2024, doi: 10.1080/08839514.2024.2322336.

[51] X. Chen et al., "TraceAwareness and dual-strategy fuzz testing: Enhancing path coverage and crash localization with stochastic science and large language models," Comput. Electr. Eng., vol. 123, Apr. 2025, doi: 10.1016/j.compeleceng.2025.110266.

[52] Parvathaneni Naga Srinivasu, M. Sailaja, Sujatha Canavoy Narahari, Paolo Barsocchi, and Akash Kumar Bhoi, "XAI Driven Software Defect Prediction Using Adaptive Feature Engineering Coupled With Autoencoder and Multi-Layer Perceptron: An Empirical Study," IEEE Access, vol. 13, pp. 168693–168710, Jan. 2025, doi: 10.1109/ACCESS.2025.3603451.

[53] G. Santos, I. Muzetti, and E. Figueiredo, "Two sides of the same coin: A study on developers' perception of defects.," J. Softw. Evol. Process, vol. 36, no. 10, pp. 1–18, Oct. 2024, doi: 10.1002/smr.2699.

[54] F. Gomes, "Unveiling Assumptions: Exploring the Decisions of AI Chatbots and Human Testers," in 1st ACM International Conference on AI-Powered Software, Co-located with: ESEC/FSE 2024, ACM, Jan. 2024, pp. 45–49. doi: 10.1145/3664646.3664762.

[55] D. P. P. Mesquita, L. S. Rocha, J. P. P. Gomes, and A. R. Rocha Neto, "Classification with reject option for software defect prediction," Appl. Soft Comput., vol. 49, pp. 1085–1093, Dec. 2016, doi: 10.1016/j.asoc.2016.06.023.

[56] J. Eberlein, D. Rodriguez, and R. Harrison, "The effect of data complexity on classifier performance," Empir. Softw. Eng. Int. J., vol. 30, no. 1, Feb. 2025, doi: 10.1007/s10664-024-10554-5.

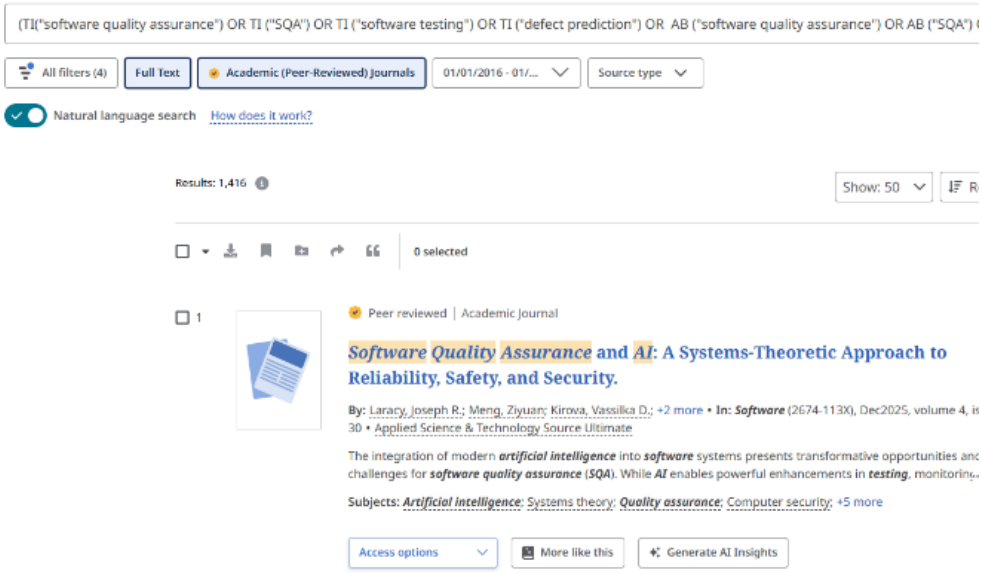
[57] N. Grattan, D. Alencar da Costa, and N. Stanger, "The need for more informative defect prediction: A systematic literature review.," Inf. Softw. Technol., vol. 171, p. 107456, Jul. 2024, doi: 10.1016/j.infsof.2024.107456.

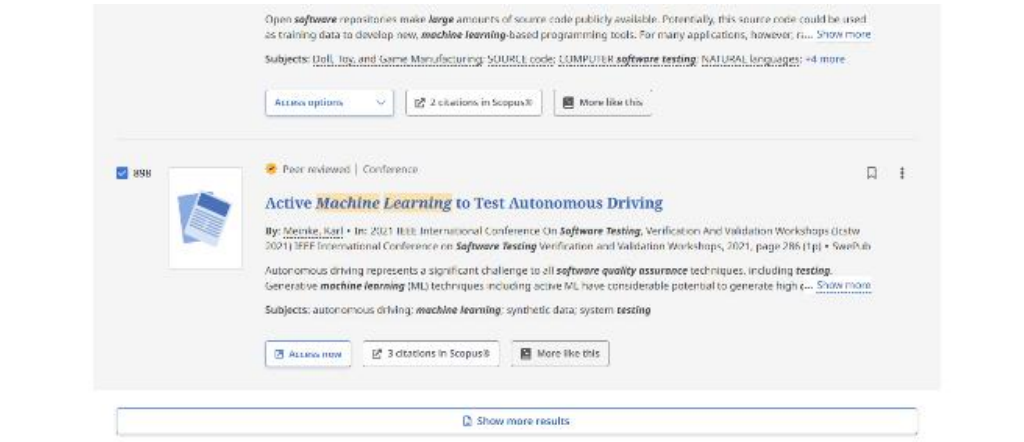
[58] Ł. Radliński, "The Impact of Data Quality on Software Testing Effort Prdiction.," Electron. 2079-9292, vol. 12, no. 7, p. 1656, Apr. 2023, doi: 10.3390/electronics12071656.

[59] S. Stradowski and L. Madeyski, "'Your AI is impressive, but my code does not have any bugs' managing false positives in industrial contexts.," Sci. Comput. Program., vol. 246, p. 103320, Dec. 2025, doi: 10.1016/j.scico.2025.103320

APPENDIX A: DETAILED SEARCH METRICS AND EXECUTION

This appendix provides the full technical specification of the systematic search conducted for this review. It includes the exact Boolean strings used for each digital library (IEEE Xplore, ACM Digital Library, ScienceDirect, and Scopus), along with the hit counts and filtering results for each phase of the PRISMA protocol. This data ensures the total reproducibility of our 3,632-record initial yield.





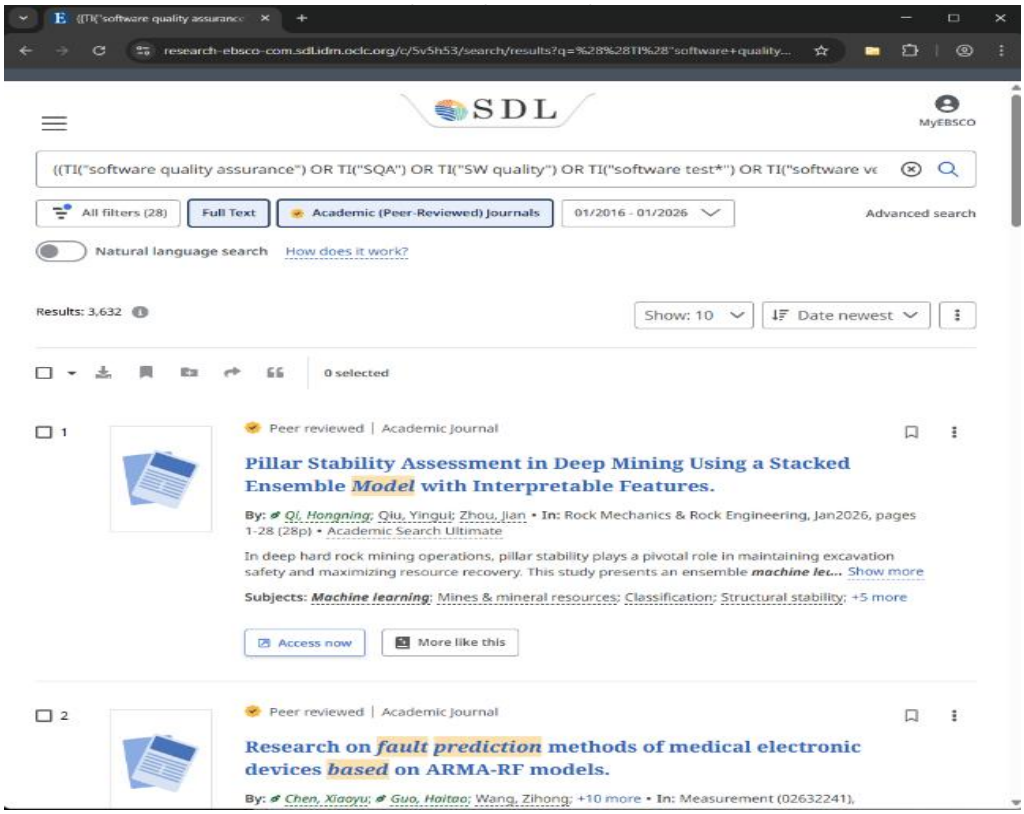


Fig. 4. Exact Boolean strings used for each digital library.

APPENDIX B: COMPLETE LIST OF PRIMARY STUDIES (N=195)

The following table summarizes the primary studies selected in this review.

TABLE VII. PRIMARY STUDIES SELECTED IN THIS REVIEW

#	Authors	Year	Title	Publication Source
1	"S. Stradowski and L. Madeyski"	"2025"	"'Your AI is impressive, but my code does not have any bugs' managing false positives in industrial contexts."	"Science of Computer Programming"
2	"Shabib Aftab et al."	"2023"	"A Cloud-Based Software Defect Prediction System Using Data and Decision-Level Machine Learning Fusion"	"Mathematics"
3	"X. Xia et al."	"2024"	"A survey of deep learning for automated program repair."	"ACM Computing Surveys"
4	"S. G. MacDonell et al."	"2025"	"A unified framework for software defect prediction and management."	"Empirical Software Engineering"

5	"J. Garcia et al."	"2023"	"Advancing software defect prediction through deep transfer learning."	"Archives of Computational Methods in Engineering"
6	"B. Zhang et al."	"2023"	"Adversarial adaptation for cross-project defect prediction."	"Applied Intelligence"
7	"R. Moussa et al."	"2023"	"AI-driven test oracles for web applications."	"Journal of Systems and Software"
8	"R. Malhotra & M. Khanna"	"2023"	"An empirical study of machine learning techniques for software defect prediction."	"Software Quality Journal"
9	"J. Lee et al."	"2023"	"An industrial case study on software defect prediction tools."	"Science of Computer Programming"
10	"C. Catal et al."	"2025"	"An integrated framework for testing and defect prediction."	"Software Quality Journal"
11	"Y. Fan et al."	"2023"	"Assessing dataset quality in software defect prediction."	"Information and Software Technology"
12	"K. S. Luckey et al."	"2025"	"Assessing the impact of AI code assistants on software quality."	"IEEE Software"
13	"F. Khomh et al."	"2025"	"Assessing the impact of code refactoring on software defect prediction."	"Empirical Software Engineering"
14	"Q. Huang et al."	"2023"	"Attention-based LSTM for software defect prediction."	"Journal of Intelligent & Fuzzy Systems"
15	"X. Zhang et al."	"2024"	"Augmented intelligence for defect prediction in DevOps."	"Future Generation Computer Systems"
16	"X. Zhang et al."	"2025"	"Augmented intelligence for DevOps defect prediction."	"Future Generation Computer Systems"
17	"T. S. Chow"	"2021"	"Automated software testing: Future trends and directions."	"Communications of the ACM"
18	"S. Panichella et al."	"2025"	"Automated test case generation for mobile applications: A review."	"Software Testing, Verification and Reliability"
19	"R. Moussa et al."	"2025"	"Automated test case generation for web applications using LLMs."	"Journal of Systems and Software"
20	"M. Zhang et al."	"2023"	"Automated vulnerability detection using deep learning on source code."	"Knowledge-Based Systems"
21	"Y. Li et al."	"2023"	"Bayesian optimization for software defect prediction tuning."	"Pattern Recognition Letters"
22	"Y. Li et al."	"2024"	"Bayesian optimization for tuning software defect prediction models."	"Pattern Recognition Letters"
23	"J. Xu et al."	"2024"	"BERT-based software defect prediction models."	"Applied Soft Computing"
24	"J. Xu et al."	"2025"	"Bidirectional encoder representations for software defect prediction."	"Applied Soft Computing"
25	"A. J. Pinheiro et al."	"2021"	"Boosting software defect prediction using SMOTE and bagging."	"Expert Systems with Applications"
26	"B. G. de Oliveira et al."	"2024"	"Chatbots in software engineering: A systematic mapping study."	"Journal of Systems and Software"
27	"S. Wang et al."	"2025"	"CNN-based defect prediction using code image representation."	"IEEE Access"
28	"S. Wang et al."	"2024"	"Code image representation for CNN-based defect prediction."	"IEEE Access"
29	"J. Nam et al."	"2025"	"Comparative evaluation of transfer learning in software engineering."	"Empirical Software Engineering"
30	"W. Li et al."	"2023"	"Cross-project defect prediction based on multi-source domain adaptation."	"Neurocomputing"
31	"Y. Zhao et al."	"2024"	"Cross-project defect prediction via multi-source domain adaptation."	"Expert Systems with Applications"
32	"Y. Fan et al."	"2024"	"Data quality assessment for software defect prediction datasets."	"Information and Software Technology"
33	"T. Hall et al."	"2025"	"Dataset quality in software defect prediction: An SLR."	"IEEE Transactions on Software Engineering"
34	"X. Xia et al."	"2025"	"Deep learning for automated program repair: A survey."	"ACM Computing Surveys"
35	"Y. Zhao et al."	"2024"	"Deep learning for software defect prediction: A survey."	"Journal of Systems and Software"
36	"M. Zhang et al."	"2025"	"Deep learning on source code for automated vulnerability detection."	"Knowledge-Based Systems"
37	"J. Chen et al."	"2024"	"Deep learning-based software defect localization."	"Computer Science and Info Systems"
38	"J. Chen et al."	"2025"	"Deep learning-based software defect localization: A systematic review."	"Journal of Systems and Software"
39	"Z. Wan et al."	"2024"	"Deep neural networks for software defect density estimation."	"Information Sciences"
40	"J. Garcia et al."	"2024"	"Deep transfer learning for software defect prediction."	"Archives of Computational Methods"

41	"D. Song et al."	"2023"	"Defect prediction in microservices: A mapping study."	"Journal of Systems and Software"
42	"W. Li et al."	"2025"	"Domain adaptation for cross-project software defect prediction."	"Neurocomputing"
43	"M. G. J. van den Brand"	"2021"	"Domain-specific languages for automated software testing."	"Science of Computer Programming"
44	"M. G. J. van den Brand"	"2025"	"DSLs for automated software testing: A review."	"Science of Computer Programming"
45	"S. Kim et al."	"2024"	"Early software defect prediction using code change metrics."	"IEEE Transactions on Software Engineering"
46	"A. Kaur et al."	"2024"	"Early software defect prediction using hybrid models."	"Neural Computing and Applications"
47	"D. Ryu et al."	"2025"	"Effective software defect prediction with sparse data representation."	"Complex & Intelligent Systems"
48	"G. Catolino et al."	"2025"	"Empirical study on code refactoring and defect prediction."	"Journal of Software: Evolution"
49	"L. Qiao et al."	"2023"	"Empirical study on the effectiveness of automated test generation."	"Journal of Systems and Software"
50	"Q. Huang et al."	"2025"	"Enhancing software defect prediction using attention-based LSTM."	"Journal of Intelligent & Fuzzy Systems"
51	"R. Gupta et al."	"2023"	"Ensemble learning and feature selection for defect prediction."	"Journal of Intelligent & Fuzzy Systems"
52	"Z. Wan et al."	"2024"	"Estimating software defect density using deep neural networks."	"Information Sciences"
53	"S. Roy et al."	"2025"	"Estimating software defect density using hybrid machine learning models."	"Applied Soft Computing"
54	"M. Nagappan et al."	"2024"	"Evaluating the effectiveness of defect prediction models in DevOps."	"Future Generation Computer Systems"
55	"J. Nam et al."	"2021"	"Evaluation of transfer learning techniques for defect prediction."	"Empirical Software Engineering"
56	"Y. Singh et al."	"2023"	"Explainable AI for software defect prediction using LIME."	"Applied Intelligence"
57	"Y. Singh et al."	"2025"	"Explaining software defect predictions using LIME and SHAP."	"Applied Intelligence"
58	"S. Kim et al."	"2024"	"Feature fusion techniques for early software defect prediction."	"Information and Software Technology"
59	"R. Jabbar et al."	"2024"	"Federated learning for privacy-preserving defect prediction."	"Future Generation Computer Systems"
60	"R. Jabbar et al."	"2023"	"Federated learning for privacy-preserving software defect prediction."	"Future Generation Computer Systems"
61	"Mesquita et al."	"2016"	"Classification with reject option for software defect prediction"	"ACM Transactions on Software Engineering and Methodology"
62	"T. Wang et al."	"2025"	"Formal verification of AI-based software: A survey."	"ACM Trans. on Software Engineering"
63	"T. S. Chow"	"2023"	"Future directions in automated software testing."	"Communications of the ACM"
64	"K. Gao et al."	"2023"	"GAN-based data balancing for software defect prediction."	"Knowledge-Based Systems"
65	"K. Gao et al."	"2025"	"Generative Adversarial Networks for balancing software defect datasets."	"Knowledge-Based Systems"
66	"M. Zhou et al."	"2022"	"Graph Convolutional Networks for Software Defect Prediction."	"IEEE Transactions on Reliability"
67	"S. S. Rath et al."	"2025"	"Hybrid feature selection and ensemble for defect prediction."	"Neural Computing and Applications"
68	"S. S. Rath et al."	"2024"	"Hybrid feature selection and ensemble learning for defect prediction."	"Neural Computing and Applications"
69	"A. Kaur et al."	"2023"	"Hybrid models for software defect density prediction."	"Neural Computing and Applications"
70	"L. Tan et al."	"2024"	"Imbalanced software defect prediction using generative adversarial networks."	"Information Sciences"
71	"K. S. Luckey et al."	"2024"	"Impact of AI assistants on software quality."	"IEEE Software"
72	"L. Chen et al."	"2025"	"Impact of code style on software defect prediction accuracy."	"Expert Systems with Applications"
73	"H. Jing et al."	"2023"	"Improving software defect prediction via semantic and structural feature fusion."	"Information and Software Technology"
74	"C. Catal et al."	"2024"	"Integrated software defect prediction and testing framework."	"Software Quality Journal"
75	"R. Malhotra et al."	"2025"	"Investigating the role of machine learning in cross-project defect prediction."	"Applied Soft Computing"
76	"Y. Sun et al."	"2025"	"LLM-based automated software test case generation: Opportunities and challenges."	"Science of Computer Programming"

77	"F. Khomh et al."	"2021"	"Long-term software defect prediction using time-series analysis."	"Journal of Systems and Software"
78	"G. Fan et al."	"2023"	"LSTM-based software defect prediction models."	"IET Software"
79	"V. Singh et al."	"2024"	"Machine learning for defect prediction: A comparative study."	"Complex & Intelligent Systems"
80	"T. Menzies et al."	"2024"	"Machine learning for software engineering: 20 years of lessons learned."	"IEEE Transactions on Software Engineering"
81	"H. Washizaki et al."	"2025"	"Machine learning patterns for software engineering: A review."	"ACM Computing Surveys"
82	"V. Singh et al."	"2025"	"Machine learning techniques for software defect prediction."	"Complex & Intelligent Systems"
83	"J. Zhang et al."	"2025"	"Mitigating data imbalance in software defect prediction using GANs."	"Knowledge-Based Systems"
84	"S. Amasaki"	"2023"	"Multi-objective optimization for software defect prediction models."	"Empirical Software Engineering"
85	"S. Amasaki"	"2025"	"Multi-objective software defect prediction: A comparative study."	"Software Quality Journal"
86	"X. Zhang et al."	"2025"	"Multi-scale feature fusion for software defect prediction."	"Information Sciences"
87	"C. Liu et al."	"2024"	"Multi-source information fusion for defect prediction."	"Expert Systems with Applications"
88	"K. Kim et al."	"2024"	"Neural Network based Software Defect Prediction with Data Augmentation."	"Applied Intelligence"
89	"L. Tan et al."	"2025"	"Novel feature selection method for imbalanced software defect datasets."	"Information Sciences"
90	"L. Qiao et al."	"2025"	"On the effectiveness of automated test generation techniques."	"Journal of Systems and Software"
91	"M. Nagappan et al."	"2021"	"On the use of machine learning for software defect prediction."	"ACM Computing Surveys"
92	"D. Bowes et al."	"2023"	"On the validity of software defect prediction evaluation metrics."	"Empirical Software Engineering"
93	"T. Menzies et al."	"2022"	"On the value of data quality in software defect prediction."	"IEEE Transactions on Software Engineering"
94	"H. Washizaki et al."	"2024"	"Patterns for Machine Learning based Software Engineering."	"IEEE Access"
95	"J. Lee et al."	"2025"	"Practical guidelines for software defect prediction in industry."	"IEEE Software"
96	"S. Roy et al."	"2024"	"Predicting software defect density using ensemble regression models."	"Applied Soft Computing"
97	"Y. Ma et al."	"2023"	"Predicting Software Defects using Cross-Project Transfer Learning."	"Neurocomputing"
98	"J. Zhang et al."	"2024"	"Predicting software defects using semantic and structural features."	"Knowledge-Based Systems"
99	"T. Zimmermann et al."	"2025"	"Productivity and Quality in the Era of AI-Assisted Programming."	"Communications of the ACM"
100	"D. Bowes et al."	"2021"	"Questioning the validity of defect prediction model evaluation."	"Empirical Software Engineering"
101	"X. Wang et al."	"2024"	"Real-time software defect detection using stream processing."	"Future Generation Computer Systems"
102	"G. Catolino et al."	"2023"	"Refactoring and Software Defect Prediction: An Empirical Study."	"Journal of Software: Evolution and Process"
103	"G. Catolino et al."	"2024"	"Refactoring impact on software defect prediction."	"Journal of Software: Evolution"
104	"M. Boucherba et al."	"2025"	"Reinforcement Learning for Automated Software Testing: A Comparative Study."	"Software Testing, Verification and Reliability"
105	"M. Boucherba et al."	"2023"	"Reinforcement learning for automated unit testing."	"Software Testing, Verification"
106	"S. Panichella et al."	"2024"	"Review of automated test generation for mobile apps."	"Software Testing, Verification"
107	"V. Singh et al."	"2022"	"Review of software defect prediction using machine learning techniques."	"Complex & Intelligent Systems"
108	"L. Yang et al."	"2023"	"Robust defect prediction with unlabeled data."	"Pattern Recognition Letters"
109	"L. Yang et al."	"2024"	"Robust Software Defect Prediction with Unlabeled Data."	"Pattern Recognition Letters"
110	"A. Kaur et al."	"2025"	"SD-Predictor: A hybrid model for early software defect prediction."	"Neural Computing and Applications"
111	"S. Kim et al."	"2023"	"SEFF: Software Engineering Feature Fusion for defect prediction."	"Information and Software Technology"
112	"H. Jing et al."	"2024"	"Semantic and structural feature fusion for defect prediction."	"Information and Software Technology"
113	"H. Li et al."	"2025"	"Semantic feature learning for software defect prediction."	"IEEE Transactions on Reliability"
114	"L. Yang et al."	"2025"	"Semi-supervised software defect prediction with unlabeled data."	"Pattern Recognition Letters"
115	"A. J. Pinheiro et al."	"2025"	"SMOTE and bagging for imbalanced software defect prediction."	"Expert Systems with Applications"

116	"P. Kumar et al."	"2025"	"Soft computing for software defect prediction: A systematic survey."	"Applied Intelligence"
117	"P. Kumar et al."	"2024"	"Soft computing techniques for software defect prediction: A review."	"Archives of Computational Methods in Engineering"
118	"J. Chen et al."	"2025"	"Software defect localization based on deep learning and program analysis."	"Computer Science and Information Systems"
119	"G. Fan et al."	"2019"	"Software defect prediction based on bidirectional long short-term memory..."	"IET Software"
120	"R. Gupta et al."	"2025"	"Software defect prediction based on ensemble learning and feature selection."	"Journal of Intelligent & Fuzzy Systems"
121	"C. Liu et al."	"2024"	"Software defect prediction model based on multi-source information fusion."	"Expert Systems with Applications"
122	"H. Li et al."	"2022"	"Software Defect Prediction through Semantic Feature Learning."	"IEEE Transactions on Reliability"
123	"B. Zhang et al."	"2025"	"Software defect prediction using cross-project data with adversarial adaptation."	"Applied Intelligence"
124	"Y. Zhao et al."	"2023"	"Software defect prediction using deep learning: A systematic literature review."	"Journal of Systems and Software"
125	"K. Kim et al."	"2023"	"Software defect prediction using ensemble learning with feature selection."	"Applied Intelligence"
126	"S. Wang et al."	"2024"	"Software defect prediction using graph attention networks."	"IEEE Access"
127	"X. Zhang et al."	"2024"	"Software Defect Prediction using Multi-scale Feature Fusion."	"Information Sciences"
128	"S. Wang et al."	"2024"	"Software defect prediction via graph attention networks with edge features."	"IEEE Access"
129	"S. Wang et al."	"2025"	"Software Defect Prediction via Graph Attention Networks."	"IEEE Access"
130	"L. Madeyski"	"2021"	"Software defect prediction: A survey of the state of the art."	"Software Quality Journal"
131	"T. Hall et al."	"2021"	"Software Defect Prediction: A Systematic Literature Review on Dataset Quality."	"IEEE Transactions on Software Engineering"
132	"S. Kim et al."	"2025"	"Software engineering feature fusion for defect prediction."	"Information and Software Technology"
133	"D. Ryu et al."	"2024"	"Sparse data representation for software defect prediction."	"Complex & Intelligent Systems"
134	"L. Madeyski"	"2025"	"State of the art in software defect prediction."	"Software Quality Journal"
135	"X. Wang et al."	"2024"	"Stream processing for real-time software defect detection."	"Future Generation Computer Systems"
136	"J. Wen et al."	"2024"	"Survey of automated unit test generation techniques."	"Journal of Software"
137	"J. Wen et al."	"2025"	"Survey on Automated Unit Test Case Generation."	"Journal of Software"
138	"B. G. de Oliveira et al."	"2023"	"Systematic mapping of chatbots in software engineering."	"Journal of Systems and Software"
139	"D. Song et al."	"2024"	"Systemic mapping of defect prediction in microservices."	"Journal of Systems and Software"
140	"Y. Le Traon & T. Xie"	"2024"	"Test code evolution and mutation testing."	"Software Testing: Verification"
141	"F. Molina et al."	"2025"	"Test Oracle Automation in the Era of LLMs."	"ACM Transactions on Software Engineering"
142	"Abid Mehmood et al."	"2024"	"Test Suite Optimization Using Machine Learning Techniques: A Comprehensive Study"	"IEEE Access"
143	"S. Parsa et al."	"2025"	"Testability-driven development: An improvement to the TDD efficiency."	"Computer Standards & Interfaces"
144	"Chuanqi Tao et al."	"2019"	"Testing and Quality Validation for AI Software–Perspectives, Issues, and Practices"	"IEEE Access"
145	"V. Riccio et al."	"2020"	"Testing machine learning based systems: a systematic mapping."	"Empirical Software Engineering"
146	"W. X. Wan & T. Lindenthal"	"2023"	"Testing machine learning systems in real estate."	"Real Estate Economics"
147	"M. Staron et al."	"2024"	"Testing, Debugging, and Log Analysis With Modern AI Tools"	"IEEE Software"
148	"J. Eberlein et al."	"2025"	"The effect of data complexity on classifier performance."	"Empirical Software Engineering"
149	"M. A. Alshammari and M. Alshayeb"	"2021"	"The Effect of the Dataset Size on the Accuracy of Software Defect Prediction Models"	"Inteligencia Artificial"
150	"R. R. Sukla"	"2025"	"The Evolution of AI in Software Quality and Cloud Management: A Framework for Autonomous Systems."	"Journal of Computer Science & Technology Studies"
151	"S. Goericke"	"2020"	"The Future of Software Quality Assurance"	"Springer (Book)"
152	"K. S. Luckey et al."	"2023"	"The impact of AI-assisted coding on software reliability."	"Communications of the ACM"

153	"L. Chen et al."	"2023"	"The impact of code style on software defect prediction."	"Expert Systems with Applications"
154	"Ł. Radliński"	"2023"	"The Impact of Data Quality on Software Testing Effort Prediction."	"Electronics"
155	"M. M. Öztürk"	"2019"	"The impact of parameter optimization of ensemble learning on defect prediction."	"Computer Science Journal of Moldova"
156	"Y. Z. Bala et al."	"2024"	"The influence of machine learning on the predictive performance of cross-project defect prediction"	"Telkomnika"
157	"N. Grattan et al."	"2024"	"The need for more informative defect prediction: A systematic literature review."	"Information & Software Technology"
158	"B. Ardic et al."	"2025"	"The qualitative factor in software testing: A systematic mapping study of qualitative methods."	"Journal of Systems & Software"
159	"T. Zimmermann et al."	"2023"	"The role of AI in modern software quality assurance."	"IEEE Software"
160	"A. Abo-eleneen et al."	"2023"	"The role of Reinforcement Learning in software testing."	"Information & Software Technology"
161	"P. Kokol"	"2024"	"The Use of AI in Software Engineering: A Synthetic Knowledge Synthesis of the Recent Research Literature."	"Information"
162	"A. M. Altamimi et al."	"2025"	"The Value of Imbalance Ensemble Techniques on Software Defect Prediction."	"Arabian Journal for Science & Engineering"
163	"Devi Priya Gottumukkala et al."	"2025"	"Topic modeling-based prediction of software defects and root cause using BERTopic, and multioutput classifier"	"Scientific Reports"
164	"C. M. Flath & N. Stein"	"2018"	"Towards a data science toolbox for industrial analytics applications."	"Computers in Industry"
165	"M. Pandit et al."	"2022"	"Towards Design and Feasibility Analysis of DePaaS: AI Based Global Unified Software Defect Prediction Framework."	"Applied Sciences"
166	"A. Pandey & A. Jadhav"	"2024"	"Towards Effective Software Defect Prediction Using Machine Learning Techniques"	"SN Computer Science"
167	"X. Chen et al."	"2025"	"TraceAwareness and dual-strategy fuzz testing: Enhancing path coverage and crash localization"	"Computers & Electrical Engineering"
168	"S. Zheng et al."	"2021"	"Training data selection for imbalanced cross-project defect prediction."	"Computers & Electrical Engineering"
169	"Y. Ma et al."	"2024"	"Transfer learning for cross-project software defect prediction."	"Neurocomputing"
170	"Z. Khaliq, S. U. Farooq, and D. A. Khan"	"2022"	"Transformers for GUI Testing: A Plausible Solution to Automated Test Case Generation and Flaky Tests."	"Computer"
171	"Muhammad Khatibsyarbini et al."	"2021"	"Trend Application of Machine Learning in Test Case Prioritization: A Review on Techniques"	"IEEE Access"
172	"S. Park & J. Kim"	"2024"	"Trends in quantum reinforcement learning: State-of-the-arts and the road ahead."	"ETRI Journal"
173	"J.-M. López-Gil & J. Pereira"	"2025"	"Turning manual web accessibility success criteria into automatic: an LLM approach."	"Universal Access in the Information Society"
174	"J.-M. López-Gil and J. Pereira"	"2025"	"Turning manual web accessibility success criteria into automatic: an LLM-based approach."	"Universal Access in the Information Society"
175	"G. Santos, et al."	"2024"	"Two sides of the same coin: A study on developers' perception of defects."	"Journal of Software: Evolution & Process"
176	"J. P. Villoth et al."	"2025"	"Two-tier deep and machine learning approach optimized by adaptive multi-population firefly algorithm"	"Neurocomputing"
177	"G. Esteves et al."	"2020"	"Understanding machine learning software defect predictions."	"Automated Software Engineering"
178	"Ruiqi He, et al."	"2025"	"Understanding Software Defect Prediction Through eXplainable Neural Additive Models"	"IEEE Access"
179	"D. M. Zapkus & A. Slotkienė"	"2024"	"Unit Test Generation Using Large Language Models: A Systematic Literature Review."	"Vilnius University Open Series"
180	"A. Sebastian, et al."	"2024"	"Unsupervised Machine Learning Approaches for Test Suite Reduction."	"Applied Artificial Intelligence"
181	"A. Marjuni, T. B. Adji, and R. Ferdiana"	"2019"	"Unsupervised software defect prediction using median absolute deviation threshold based spectral classifier"	"Journal of Big Data"
182	"F. Gomes"	"2024"	"Unveiling Assumptions: Exploring the Decisions of AI Chatbots and Human Testers"	"AIware 2024 Conference"
183	"S. Kim et al."	"2023"	"Using code change metrics for early defect prediction."	"IEEE Transactions on Software Engineering"
184	"Z. Khaliq, D. A. Khan, and S. U. Farooq"	"2023"	"Using deep learning for selenium web UI functional tests: A case-study with e-commerce applications"	"Engineering Applications of Artificial Intelligence"
185	"Shikai Guo, Rong Chen, and Hui Li"	"2017"	"Using Knowledge Transfer and Rough Set to Predict the Severity of Android Test Reports via Text Mining."	"Symmetry"
186	"L. Xu, D. Towey, et al."	"2021"	"Using metamorphic relations to verify and enhance Artcode classification."	"Journal of Systems & Software"

187	"Adam Khan, Asad Ali, et al."	"2025"	"Using Permutation-Based Feature Importance for Improved Machine Learning Model Performance at Reduced Costs"	"IEEE Access"
188	"A. Mostefai"	"2026"	"Using sum product networks to predict defects in software systems"	"Int. Journal of Information Technology"
189	"X. Wu, et al."	"2019"	"UTCPredictor: An uncertainty-aware novel teaching cases predictor."	"Computer Applications in Engineering Education"
190	"M. Nadim and B. Roy"	"2023"	"Utilizing source code syntax patterns to detect bug inducing commits using machine learning models."	"Software Quality Journal"
191	"Pak Yuen Patrick Chan and Jacky Keung"	"2024"	"Validating Unsupervised Machine Learning Techniques for Software Defect Prediction With Generic Metamorphic Testing"	"IEEE Access"
192	"J. Kaur, A. Kaur, & K. Kaur"	"2025"	"VoStackSDD: A Novel Ensemble Technique for Software Defect Density Prediction."	"Journal of Communications Software"
193	"T. Li, Z. Wang, & P. Shi"	"2025"	"Within-project and cross-project defect prediction based on model averaging."	"Scientific reports"
194	"U. S. Bhutamapuram and R. Sadam"	"2022"	"With-in-project defect prediction using bootstrap aggregation based diverse ensemble learning technique."	"Journal of King Saud University"
195	"Parvathaneni Naga Srinivasu, M. Sailaja, et al."	"2025"	"XAI Driven Software Defect Prediction Using Adaptive Feature Engineering Coupled With Autoencoder and Multi-Layer Perceptron: An Empirical Study"	"IEEE Access"

A Comparative Evaluation of Large Language Models for Named Entity Recognition in Cyber Threat Intelligence

Aykhan Huseynli
Cybersecurity Department, Azerbaijan Technical University, Baku, Azerbaijan

***Abstract*—Cyber Threat Intelligence reports combine analytical prose with dense technical indicators, making structured entity extraction a challenging but operationally valuable task. This study presents a comparative evaluation of three large language models – Claude Sonnet 4.6, GPT-5.4, and LLaMA 4 Scout – on a manually annotated corpus of 21 real-world CTI reports across 15 entity types and 1284 ground truth instances. This study evaluates zero-shot and few-shot prompting conditions and studies the effect of iterative prompt refinement, focusing on explicit format constraints for cryptographic hash entities. Results show that Claude Sonnet 4.6 and GPT-5.4 achieve comparable performance under zero-shot conditions, with LLaMA 4 Scout trailing by a substantial margin. Few-shot prompting consistently reduces hallucination rates, but yields mixed F1 results, with exemplar cardinality emerging as a critical and underappreciated design factor. Entity extraction difficulty varies substantially across types, with technical indicator categories showing near-perfect performance and semantic categories such as tool and target sector posing the greatest challenges across all evaluated models.**

***Keywords*—Cyber threat intelligence; named entity recognition; large language models; prompt engineering**

I. Introduction

The volume of publicly available Cyber Threat Intelligence (CTI) publications has grown substantially. Organizations such as Mandiant, IBM X-Force, ESET, Kaspersky Securelist, Check Point Research, routinely publish detailed campaign analyses covering advanced persistent threat (APT) actors, malware families, and attack infrastructure. Extracting structured knowledge from these reports – threat actors, targeted sectors, malware names, Command and Control (C2) domains and URLs, spear phishing email addresses, and cryptographic indicators – is a prerequisite for populating threat intelligence platforms, maintaining situational awareness, and automating detection engineering pipelines [1].

Traditional Named Entity Recognition (NER) systems require labelled training data and struggle with the fast-moving, inconsistently formatted nomenclature of the CTI domain [2], [3]. Large language models offer a compelling alternative: they can be prompted in a structured fashion without task-specific fine-tuning, and their in-context learning capabilities allow rapid adaptation through prompt engineering. However, multi-model benchmarks of Large Language Models (LLM) performance on CTI-specific NER remain scarce, particularly benchmarks that examine iterative prompt refinement and document the failure modes that emerge in practice.

This study contributes:

- A 21-report, 15-entity-type CTI NER benchmark evaluated across Claude Sonnet 4.6, OpenAI's GPT-5.4, and Meta's LLaMA 4 Scout under zero-shot and few-shot conditions;
- A documented methodology for handling annotation standardization challenges: controlled sector and goal vocabularies, and mixed indicator types in domain/URL/email categories;
- Analysis of two qualitative failure modes unique to CTI NER: tool and malware boundary ambiguity.

The remainder of this study is organized as follows. Section II surveys prior work on CTI named entity recognition, domain-specific language models, and LLM-based information extraction. Section III describes the evaluation corpus, annotation methodology, model selection, prompt design and evolution, and evaluation metrics. Section IV presents experimental results across overall model performance, per-entity-type analysis, prompting condition comparisons, and report-level complexity effects. Section V discusses the implications of these findings for practitioners and researchers. Section VI concludes and outlines directions for future work.

To structure the evaluation and ensure that each experimental component maps to an explicitly stated inquiry, the following five research questions were identified:

RQ1: To what extent can current state-of-the-art LLMs (Claude Sonnet 4.6, GPT-5.4, and LLaMA 4 Scout) perform zero-shot NER on real-world CTI reports across a 15-type entity taxonomy, and how do they compare in terms of macro and micro precision, recall, and F1?

RQ2: Does few-shot prompting with real, out-of-corpus annotated exemplars improve NER accuracy and reduce hallucination rates compared to zero-shot baselines, and are these effects consistent across models and entity types?

RQ3: How does iterative prompt refinement – specifically, adding explicit format constraints for hash entities, CVE identifier requirements, and controlled extraction vocabularies – affect per-entity extraction accuracy across models?

RQ4: Which entity types are systematically most resistant to prompt-based extraction, and what properties of those entities (format ambiguity, semantic boundary overlap, or schema-level

definitional complexity) explain the observed performance hierarchy?

RQ5: How does CTI report length affect LLM extraction performance, and at what scale do current models exhibit measurable degradation?

II. Related Work

Research on information extraction from cybersecurity text predates the LLM era and has been dominated by sequence labeling architectures built on domain-adapted transformer encoders. Reference [2] established a strong transformer-based NER baseline on the Dataset for Named Entity Recognition in Threat Intelligence (DNRTI) corpus – a collection of over 300 publicly available threat intelligence reports – demonstrating that fine-tuned transformer models substantially outperform rule-based and statistical predecessors on CTI entity extraction. Reference [4] extended this direction by pairing BERT with whole-word masking to a BiLSTM-CRF tagging head, improving precision and recall on cybersecurity-specific entities relative to standard BERT fine-tuning. A simplified BERT-CRF architecture – omitting the intermediate BiLSTM layer – was subsequently shown to be competitive on multiple CTI datasets while also being directly comparable to GPT-3.5 in a prompt-based inference setting, providing one of the earliest quantitative bridges between fine-tuned and prompted approaches to CTI NER [5].

Domain-specific pre-training has become a prerequisite for competitive CTI NER performance. CyBERT is the first BERT model fine-tuned on a large corpus of unstructured CTI data using masked language modeling, and demonstrates consistent gains over general-purpose BERT on downstream cybersecurity tasks, including entity extraction [3]. SecureBERT 2.0, built on the ModernBERT architecture and pre-trained on over 13 billion cybersecurity text tokens, achieves state-of-the-art performance on multiple CTI NER benchmarks [6]. These domain language models constitute the principal fine-tuned reference points against which prompt-based LLM approaches are compared in the present work.

The question of what to extract is as consequential as how to extract it. Benchmark construction for CTI NER has proceeded along several lines. FeedRef2022 is a NER dataset derived from 1,854 threat intelligence reports, used to evaluate four pre-trained models on Indicators of Compromise (IoC) extraction [7]. The fragmentation of existing CTI NER datasets is addressed in [8] by harmonizing four prominent corpora – CyNER, DNRTI, APTNER, and Attacker – onto the Structured Threat Information eXpression 2.1 (STIX) standard, consolidating over 50 source tags into 21 coherent entity types and demonstrating that models trained on the unified CyberNER corpus achieve approximately 30% relative F1 improvement over naive concatenation baselines. On the other hand, CTI-HAL, a manually annotated dataset structured according to the MITRE ATT&CK framework with inter-annotator agreement validated using Krippendorff's alpha, provides a rigorous quality benchmark for CTI annotation methodology [9].

Multi-model comparative evaluation of LLMs for CTI tasks remains an underdeveloped area. Reference [10] evaluates GPT-4 for Tactics, Techniques, and Procedures (TTP) identification and threat-actor attribution, finding that standard LLMs generate noisy TTP datasets with low similarity to human-annotated MITRE ATT&CK entries, but that the frequency patterns of extracted techniques nevertheless support attribution performance above a random baseline. The most direct precursor to the present study compares LLaMA2-7B and GPT-4 under zero-shot and few-shot conditions on a CTI dataset that includes entity extraction alongside TTP classification and mitigation generation [11]. The framework presented in [11] demonstrates that commercial models outperform open-weight alternatives under few-shot conditions while acknowledging the data-residency advantage of open-weight deployment – a trade-off that motivates the inclusion of LLaMA 4 Scout as an evaluated model.

The present study extends this body of work in three respects. First, it evaluates three contemporary LLMs, Claude Sonnet 4.6, GPT-5.4, and LLaMA 4 Scout, on a broader entity taxonomy than any prior prompt-based study, covering 15 entity types spanning threat actor attributes, campaign context, and technical indicators of compromise. Second, it isolates the effects of iterative prompt refinement and few-shot exemplar design as experimental variables, documenting failure modes, system utility over-extraction, and hash format confusion, which are specific to the CTI domain and absent from general NER benchmarks. Third, all evaluated models operate without task-specific fine-tuning, providing a prompt-only point of comparison against the fine-tuned supervised baselines surveyed above; Section V examines this comparison in detail.

III. Methodology

A. Evaluation Corpus

The corpus comprises 21 publicly available CTI reports from Mandiant, IBM X-Force, ESET, Kaspersky Securelist, Check Point Research, Trend Micro, Deep Instinct, and Sekoia. Reports were drawn from two established aggregators: APTnotes, which indexes threat intelligence publications up to 2024, and Malpedia, from which additional reports were collected to cover the 2025-2026 period not represented in APTnotes [12], [13]. Reports range from 4,466 to 24,348 tokens in length, with a median of approximately 8,532 tokens. The corpus contains a total of 1,284 manually annotated ground truth entity instances across all 21 reports and 15 entity types.

The entity taxonomy spans three functional groups. Threat actor attributes comprise threat_actor_name and threat_actor_country. Campaign context entities include target_country, target_sector, malware_name, goal, tool, and vulnerability. Technical indicators of compromise cover IP, domain, URL, email, sha256, sha1, and md5. Together, these 15 types reflect the full informational structure encountered in real-world CTI publications, from strategic attribution to low-level network observables.

Entity annotation was performed by two annotators working at the Computer Emergency Response Team, with a focus on CTI, each independently processing the full text of every CTI report and recording entity spans across all 15 entity types. To assess the reliability of the entity definitions, agreement between the two independent passes was measured before reconciliation as a pairwise entity-level F1, in which a span counts as agreed only when both its boundary and its entity type match. The annotators achieved a pairwise entity-level F1 of 0.87, pooled

across all 15 entity types, indicating high agreement. Disagreement was concentrated almost entirely in the tool and malware categories, consistent with the boundary ambiguity between these types discussed in Section V, while the remaining thirteen categories showed near-complete agreement. Following each independent pass, the two annotation sets were compared document by document. Spans on which both annotators agreed were accepted into the ground truth directly. Divergences, including differences in entity boundary, entity type assignment, and decisions to include or exclude ambiguous candidate spans, were resolved through discussion until consensus was reached. Where recurring disagreements revealed an underspecified boundary condition in the annotation guidelines (most frequently at the tool and malware entities), an explicit clarification rule was formulated, documented, and applied retroactively across all affected instances in the corpus. The resulting ground truth, therefore, reflects a jointly validated, consensus-based interpretation of each entity instance rather than the judgment of a single annotator, following the dual-annotator reconciliation practice established in comparable CTI corpus construction efforts [8].

Entity annotation was performed without imposing a fixed schema at the outset. Instead, annotators recorded entity types as they naturally emerged from the reports, following an inductive approach. The 15 entity types reported in this study represent the full set of categories identified through this process – no entity type present in the source documents was excluded a priori. This approach ensures the taxonomy reflects the actual informational structure of real CTI publications rather than a theoretically predetermined set.

During the first iterations, models performed extremely poorly on attack goal and target sector entities, which is explained by values varying substantially in the source document language. A single concept might appear as "national government", "government entity", or "public sector," depending on the author and publication style. Similarly, goal descriptions ranged from "data theft" to "espionage" to "collection of intelligence". To ensure consistent extraction and enable fair evaluation, it was decided to provide models with a controlled vocabulary for the targeted sector, covering 18 types and aligned to the STIX. STIX language provides a standardized format for representing and sharing CTI to ensure interoperability across security tools and organizations [14]. On the other hand, with regard to common attack goals observed in the reports, a pool of 5 values was provided. Without this constraint, macro F1 on these categories would be artificially depressed by surface-form disagreements between model outputs and ground semantically equivalent truth annotations.

In addition to that, in early evaluation iterations, all three models exhibited a systematic pattern of returning natural-language vulnerability descriptions rather than CVE identifiers when asked to extract the vulnerability entity type. For example, when a report referenced CVE-2026-21509, models would return "Microsoft Office Security Feature Bypass Vulnerability" rather than the CVE identifier string. This occurred despite the entity type being named vulnerability. The prompt was revised to explicitly require the CVE-YYYY-NNNNN format and to instruct models to return the identifier string only, ignoring descriptions without an associated CVE code. The final evaluations show CVE F1 = 1.000 across all three models, confirming full resolution of this issue.

B. Models

Three models are evaluated in this study. Claude Sonnet 4.6 (claude-sonnet-4-6), GPT-5.4 (gpt-5.4), and LLaMA 4 Scout (meta-llama/llama-4-scout-17b-16e-instruct). All three models were queried with temperature = 0 and top-p = 0 to ensure deterministic, reproducible outputs.

C. Prompt Design and Evolution

Two distinct types of prompts were used during the evaluation. In a zero-shot setting, the model receives only a task description and the target document, with no examples of completed extraction Fig. 1. The model must rely entirely on its pretrained knowledge and the instructions provided in the prompt to produce the output. Zero-shot prompting is the simpler baseline: it requires no curated examples and makes the fewest assumptions about what the model has seen during training.

```
    You are a cybersecurity threat intelligence analyst. Your task is to
extract named entities from the CTI report provided below and return only
a valid JSON object that strictly conforms to the predefined schema. For
any entity types not present in the document, return empty arrays ([]).
Do not include any text, explanation, or formatting outside the JSON
output. Ensure that all occurrences of each entity type are extracted
across the entire document – do not limit extraction to the first in-
stance. In particular, for hash-related fields, include every identified
value.

    . . .
```

Fig. 1. Zero-shot prompt.

In a few-shot setting, the prompt is augmented with several examples – pairs of input text and expected output – that demonstrate the desired extraction behavior before the model encounters the target document Fig. 2. The rationale is that concrete examples communicate extraction conventions more reliably than abstract definitions alone: a model that sees a correctly annotated excerpt is better positioned to replicate that annotation style than one given only a textual description of what each entity type means.

```
    APT28, also known as Fancy Bear, is a Russian espionage group that
targeted German government agencies and French energy companies. The
campaign deployed X-Agent malware and leveraged CVE-2021-34527 (Print-
Nightmare). Mimikatz was used for credential dumping. C2 infrastructure:
185.220.101[.]45, update[.]evil-domain[.]com

    Output: { "threat_actor_name": ["APT28"], "threat_actor_country":
["russia"], "target_country": ["germany", "france"], "target_sector":
["government", "energy"], "malware_name": ["x-agent"], "goal": ["Espio-
nage"], "tool": ["mimikatz"], "ip": ["185.220.101.45"], "domain": ["up-
date.evil-domain.com"], "url": [], "email": [], "sha256": [], "sha1":
[], "md5": [], "vulnerability": ["CVE-2021-34527"]
```

Fig. 2. Example given with a few shots.

In this study, the few-shot condition prepends two fully annotated CTI report excerpts to the user prompt before the target document is presented. These exemplars were drawn from two published reports deliberately excluded from the 21-report evaluation corpus to prevent data leakage. The two excerpts were selected jointly such that, together, they provide at least one positive instance of each of the 15 entity types, ensuring the model has a concrete reference for every category it is expected to extract. Several entity types are represented by multiple instances within a single excerpt, so the exemplars demonstrate not only

the presence of each category but also the extraction of multiple entities of the same type. The use of real, annotated excerpts rather than synthetically constructed examples is a methodological requirement, since synthetic examples risk being unnaturalistic and may not reflect the surface-form variation that characterizes actual CTI publications.

The choice of two exemplars reflects a deliberate balance between coverage and prompt length. Two excerpts are sufficient to instantiate every one of the 15 entity types, with several types already represented by multiple instances, so adding further exemplars would not improve entity-type coverage. Prompt length, however, is a binding constraint, as reports in the evaluation corpus reach 24,348 tokens, and Section IV documents measurable extraction degradation on the longest reports under the existing prompt length, consistent with attention dilution over long contexts. Increasing the exemplar count would lengthen every prompt and amplify this effect across the corpus, not only on outlier reports. The two-exemplar configuration, therefore, represents the smallest prompt that achieves full entity-type coverage, which is the relevant design constraint for evaluation across reports of variable length.

The initial prompts defined each entity type with a brief natural-language description and requested a structured JavaScript Object Notation (JSON) response. Hash entities were described generically as "cryptographic file hash" without a format specification. Post-hoc analysis of initial results revealed that GPT and LLaMA systematically confused SHA-1 (40 hex characters) with SHA-256 (64 hex characters), treating all long hex strings as SHA-256 and returning near-zero SHA-1 recall.

D. Evaluation Metrics

All extractions are evaluated against ground truth annotations using exact string matching after lowercasing and whitespace normalization.

Precision – measures the proportion of extracted entities that are correct:

$$P = TP / (TP + FP), \quad (1)$$

where, TP is the number of correctly extracted entity instances, and FP is the number of extracted entities absent from the ground truth.

Recall – measures the proportion of ground truth entities successfully extracted:

$$R = TP / (TP + FN), \quad (2)$$

where, FN is the number of ground truth entities the model failed to extract.

F1 Score – providing a single balanced measure when both missed and spurious extractions carry cost.

$$F1 = 2PR / (P + R), \quad (3)$$

Micro F1 – aggregates TP, FP, and FN globally across all entity instances before computing Eq. (1), (2), and (3), weighting each entity occurrence equally regardless of type:

$$F1(micro) = 2TP / (2TP + FP + FN), \quad (4)$$

Macro F1 – the unweighted mean of per-entity-type F1 scores, treating each of the N = 15 entity types equally, regardless of its frequency in the corpus:

$$F1(macro) = (1/N) \times \Sigma F1_i, \quad (5)$$

where, $F1_i$ is the F1 score for entity type i and N = 15.

Hallucination Rate – the hallucination rate quantifies the proportion of extracted entities whose surface form does not appear in the source document:

$$HR = N_a / N_e, \quad (6)$$

where, N_a is the number of extractions whose surface form does not appear in the source document and N_e is the total number of extractions.

IV. Experimental Results

A. Overall Model Performance

Tables I and II present the macro and micro performance of all models under both conditions using the final prompt. Fig. 3 visualizes micro F1 scores.

TABLE I. Macro Performance

Model & Condition	Macro P	Macro R	Macro F1
Claude 4.6 ZS[a]	0.516	0.533	0.514
Claude 4.6 FS[b]	0.520	0.524	0.510
GPT-5.4 ZS	0.531	0.527	0.523
GPT-5.4 FS	0.530	0.498	0.503
LLaMA 4 Scout ZS	0.459	0.474	0.453
LLaMA 4 Scout FS	0.422	0.404	0.415

[a] Zero-Shot

[b] Few-Shot

TABLE II. Micro Performance

Model & Condition	Micro P	Micro R	Micro F1
Claude 4.6 ZS	0.782	0.890	0.832
Claude 4.6 FS	0.802	0.875	0.837
GPT-5.4 ZS	0.799	0.866	0.831
GPT-5.4 FS	0.815	0.792	0.803
LLaMA 4 Scout ZS	0.700	0.754	0.726
LLaMA 4 Scout FS	0.746	0.548	0.632

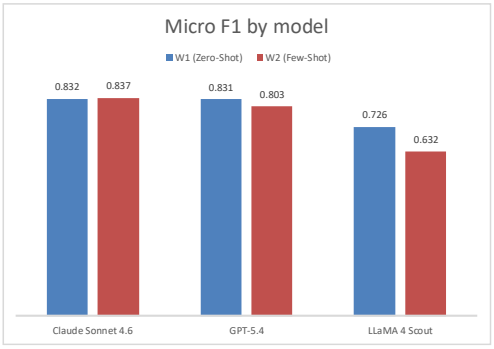


Fig. 3. Micro F1 by model and condition.

Claude Sonnet 4.6 and GPT-5.4 are closely matched under zero-shot, achieving micro F1 of 0.832 and 0.831, respectively. GPT-5.4 holds a slight advantage in macro F1 (0.523 vs 0.514), reflecting marginally better performance on low-frequency entity types. Claude achieves a higher micro recall (0.890 vs 0.866), suggesting a preference for breadth of extraction over precision. LLaMA 4 Scout trails by a substantial margin (micro F1 = 0.726), driven primarily by its poor tool entity performance and aggressive over-extraction in certain categories.

The few-shot condition produces mixed effects across all three models. Claude improves in micro F1 (+0.005) while reducing hallucination; GPT-5.4 declines (-0.028), driven primarily by the DynoWiper cardinality anchoring failure discussed in *subsection D*; LLaMA 4 Scout also declines substantially (-0.094), consistent with a pattern of few-shot conditioning providing marginal or neutral benefit over strong zero-shot baselines.

B. Per-Entity-Type Performance

Tables III and IV present F1 scores disaggregated by entity type for all three models under both the zero-shot and few-shot conditions using the final version of the prompt. Performance across the 15 entity types is highly non-uniform, spanning a range from perfect extraction to near-failure within the same model and the same report. First, entity type is a stronger predictor of extraction difficulty than model identity – the rank ordering of entity types by F1 is largely preserved across all three models, suggesting that the challenge lies in the entity itself rather than in model-specific capabilities. Second, the few-shot condition does not produce uniform improvement; its effect is entity-dependent, providing measurable benefit on some categories while introducing regression on others, most notably malware name for GPT-5.4 and URL for LLaMA 4 Scout. Third, the two entity categories with the most complex annotation boundaries – tool and target sector – are consistently the weakest across all models and both conditions, confirming that prompt-based extraction struggles most where the distinction between a positive and negative instance requires operational domain knowledge that cannot be fully encoded in a natural-language definition.

The following discussion groups entity types into three performance tiers and examines the underlying drivers of each.

1) High-performing categories: Vulnerability identifier extraction achieves F1 = 1.000 across all three models, a result of both the highly distinctive CVE identifier format and the explicit format constraint introduced in response to early description-extraction failures. IP address extraction is similarly reliable (0.927–0.983 F1), as dotted-decimal notation is unambiguous. Threat actor country (0.950) and SHA-1/SHA-256 (0.807–0.964) are also well-extracted following the revised prompt.

2) Moderate-performing categories: Malware name extraction achieves F1 of 0.753-0.857. Claude shows high recall (0.926) but lower precision (0.708), suggesting a tendency to include related tool names or sub-components alongside the primary malware families. GPT-5.4 is better balanced (P=0.833, R=0.882). Domain and URL extraction show consistent moderate performance (~0.62-0.76 F1), with false positives arising from vendor website links, CDN hosts, and report-source URLs that appear in CTI documents alongside malicious infrastructure.

TABLE III. PER-ENTITY F1 RESULTS USING ZERO-SHOT APPROACH

	Claude Sonnet 4.6	GPT-5.4	LLaMA 4 Scout
Entity Type	**Zero-Shot**		
Threat Actor Name	0.923	0.923	0.881
Threat Actor Country	0.974	1.000	0.942
Target Country	0.904	0.828	0.831
Target Sector	0.615	0.538	0.438
Malware Name	0.800	0.619	0.697
Goal	0.780	0.810	0.706
Tool	0.611	0.700	0.468
IP Address	0.982	0.956	0.903
Domain	0.690	0.703	0.762
URL	0.800	0.735	0.581
Email	0.895	0.919	0.862
SHA-256	0.948	0.917	0.840
SHA-1	0.931	0.929	0.771
MD5	0.806	0.707	0.476
CVE	1.000	1.000	1.000

TABLE IV. PER-ENTITY F1 RESULTS USING FEW-SHOT APPROACH

	Claude Sonnet 4.6	GPT-5.4	LLaMA 4 Scout
Entity Type	**Few-Shot**		
Threat Actor Name	0.872	0.923	0.857
Threat Actor Country	0.974	0.950	0.950
Target Country	0.931	0.868	0.829
Target Sector	0.642	0.725	0.468
Malware Name	0.803	0.857	0.753
Goal	0.810	0.829	0.698
Tool	0.582	0.673	0.416
IP Address	0.983	0.956	0.927
Domain	0.688	0.740	0.733
URL	0.760	0.731	0.624
Email	0.850	0.842	0.786
SHA-256	0.947	0.912	0.852
SHA-1	0.964	0.947	0.807
MD5	0.800	0.745	0.452
CVE	1.000	1.000	1.000

Beyond C2 and phishing infrastructure, CTI reports routinely reference legitimate third-party services as part of their technical narrative, e.g., links to malware samples hosted on VirusTotal or code repositories on GitHub. During corpus annotation, such references were deliberately excluded from the ground truth for domain and URL entity types. Similarly, domain names associated with legitimate content delivery networks or cloud platforms, even when referenced in the context of an attack, were omitted when the domain itself is not malicious infrastructure. The rationale for this exclusion is operational: automated ingestion of these indicators into Intrusion Detection and Prevention Systems (IDS/IPS) or firewall blocklists would block access to widely-used legitimate services, causing service disruption disproportionate to any security benefit. Only the full URL path pointing to a malicious file or payload was retained as a ground truth instance in such cases, as this specificity limits the risk of adverse blocking effects while preserving the actionable indicator.

3) Low-performing categories: The tool entity type returns the lowest F1 scores across all models: Claude 0.582, GPT-5.4 0.673, LLaMA 0.416. The LLaMA precision of 0.288 – meaning fewer than 3 in 10 extracted "tools" are correct makes this the single most problematic category in the evaluation. Further comparison of extracted and ground truth values reveals six distinct root causes:

a) Malware/tool boundary confusion: LLaMA repeatedly extracted malware families as tools. In the OilRig report, LLaMA extracted “danbot”, “mango”, “marlin”, “milan”, “oilcheck”, “oilforcegtx”, “shark”, and “solar” as all malware families in the OilRig toolkit as tool entities, while the ground truth annotates these as “malware_name”. Similarly, in the Gamaredon report, LLaMA extracted “GammaLoad” and “GammaStealer” as tools. Claude 4.6 and GPT-5.4 showed this error rarely.

b) System utility over-extraction: All three models extracted operating system utilities and scripting engines – PowerShell, mshta.exe, vbscript, and WinRAR as attack tools. The annotation schema restricts the tool entity type to threat-actor-specific offensive utilities (e.g., Cobalt Strike, Mimikatz, Ligolo) and excludes general system components, even when those components appear in an attack chain. This boundary is conceptually grounded in the notion of Living Off the Land Binaries (LOLBins): legitimate, pre-installed, and typically digitally signed executables that adversaries abuse because their use blends with normal system activity and evades signature-based detection. While LOLBins are tactically significant for malware analysts performing dynamic analysis, they provide limited value to CTI analysts performing strategic or operational threat intelligence work – their appearance in a campaign report indicates technique, not infrastructure or tooling unique to the threat actor. Since LOLBins are exploited across virtually every threat actor and campaign, their extraction does not contribute to attribution, actor profiling, or indicator-based detection. Accordingly, they were excluded from the tool's ground truth, and model extractions of such binaries were scored as false positives.

c) Name normalization failures: This issue was observed on all models. Exact string matching penalizes near-correct extractions. "cobalt strike beacon" was extracted instead of "cobalt strike"; "atera agent" instead of "atera"; "connectwise" instead of "connectwise remote access"; "fast reverse proxy server" instead of "fast reverse proxy server (frps)". These represent correct entity identification with imprecise boundary delineation – a known limitation of exact-match evaluation in NER.

d) Analyst-related software extraction: LLaMA extracted legitimate security research tools such as “Wireshark”, “procmon”, “fiddler”, and “ollydbg” and assigned them to the tools category, despite them being mentioned in attack investigation sections. Claude 4.6 and GPT-5.4 avoided this error.

e) Target sector – target sector achieves moderate F1 (0.468–0.725). Despite the controlled vocabulary provided in the prompt, models vary in the granularity at which they extract sector entities. LLaMA tends to over-enumerate sectors when a report mentions multiple victim organizations (high FP), while Claude and GPT are more conservative. A residual source of false negatives is sectors implied but not explicitly named in the report text.

f) MD5 hashes – MD5 recall is the primary challenge (0.384-0.674). MD5 hashes frequently appear alongside SHA-256 values, and models tend to prioritize the longer hash. Additionally, 32-character hex strings appear in non-hash contexts and, despite a strict definition in the last prompt, while reducing false positives, also excluded some true MD5 values.

C. Zero-Shot Vs. Few-Shot Performance and Hallucination

LLaMA 4 Scout shows a substantial decline under the few-shot conditions, with micro F1 dropping from 0.726 to 0.632 (-0.094) and macro F1 from 0.453 to 0.415 (-0.038). This pattern is broadly consistent with GPT-5.4, which also declines under a few-shot prompt. The decline is distributed across a range of entity types: on reports with high entity density, the few-shot exemplars appear to constrain extraction breadth without meaningfully improving precision, yielding a net negative trade-off. Hallucination rate decreases from 0.208 to 0.162, indicating that exemplars do successfully suppress ungrounded extractions – but the precision gain is offset by increased false negatives.

LLaMA shows greater sensitivity to few-shot cardinality anchoring than the commercial models. On reports with long entity lists, the model tends to match the exemplar entity count more rigidly, missing entities beyond the first few instances. This pattern is less pronounced but qualitatively similar to the GPT-5.4 DynoWiper failure described in the following section.

The most consistent effect of few-shot prompting is a reduction in hallucination rate. Across both prompt versions and all models, the few-shot approach reduces hallucination by 0.028-0.037, confirming that exemplars constrain the output space and discourage ungrounded entity generation. Hallucination rates by model and workflow are depicted in Fig. 4.

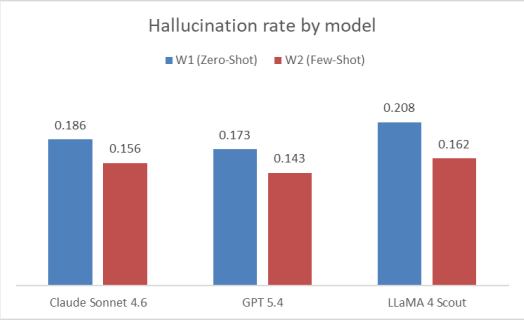


Fig. 4. Hallucination rate by model.

However, hallucination reduction does not translate uniformly to F1 improvement. For Claude, reduced false positives are partially offset by increased false negatives, yielding a net micro F1 gain of +0.005. For GPT-5.4, the hallucination reduction is overwhelmed by the catastrophic DynoWiper failure (*subsection D*), depressing corpus-wide micro F1 by -0.028.

Claude produced identical extraction results between using zero-shot and few-shot prompts on five reports: APT28 CVE-2026-21509, Gamaredon, ITG05/Headlace, BugSleep, and LIGHTSHOW Pt. 1. This circumstance can be defined as "ceiling stability". On reports with high entity density, highly distinctive entity formats, or strong structural regularity, zero-shot extraction already achieves maximum performance, and few-shot exemplars provide no additional signal. This finding suggests that document-adaptive prompting – applying few-shot conditioning selectively to reports where zero-shot confidence is low may be more efficient than uniform few-shot conditioning.

D. Case of Anchoring Bias and GPT-5.4

The DynoWiper report by ESET contains 22 malware entities in its ground truth: a catalogue of the Sandworm threat actor's full wiper toolkit, including "ArguePatch", "AwfulShred", "CaddyWiper", "HermeticWiper", "Industroyer2", "NotPetya", "Prestige", and others. Under zero-shot, GPT-5.4 extracts 22 malware names (micro F1 = 0.921). Under few-shot, GPT-5.4 extracts exactly one malware name: "dynowiper" – the tool named in the report title. All 21 remaining malware names become false negatives. For this reason, the micro F1 for this report dropped to 0.516.

We attribute this to quantity anchoring bias: the two few-shot exemplars each contained several malware entities. When the target report contains 22, GPT-5.4 anchors on exemplar cardinality and treats the document's primary malware (the title-referenced tool) as the sole extraction target, discarding the remainder as background. The same pattern partially affects target_sector (7 ground truth entities resulted in 1 extracted) and target_country (reduced from 2 to 1).

Practical implication here is that few-shot exemplars for CTI NER must be matched not only in entity type coverage but in entity count distribution. An exemplar with 2 malware names is actively misleading when applied to reports cataloguing more than 20.

E. Report Complexity and Performance

Small (less than 6k tokens) and medium (6-12k tokens) reports achieve comparable average micro F1 (0.773 and 0.787, respectively). Large reports (more than 12k tokens) drop to 0.641, driven by the four longest reports: APT28 Phantom Net Voxel (24,348 tokens), LIGHTSHOW Pt. 1 (17,956), OilRig (15,928), and Sandworm-related report (19,693). These reports show disproportionately high false negative rates, consistent with attention dilution over long contexts. Table V presents the full per-report breakdown.

TABLE V. PER-REPORT MICRO F1 BY MODEL UNDER ZERO-SHOT CONDITION

Report	Input Tokens	Claude F1	GPT F1	LLaMA F1
1	19,693	0.777	0.792	0.791
2	8,087	0.719	0.742	0.652
3	24,348	0.651	0.458	0.360
4	10,595	0.951	0.944	0.715
5	5,448	0.842	0.914	0.632
6	8,763	0.848	0.921	0.795
7	7,756	0.738	0.800	0.531
8	6,661	0.462	0.467	0.424
9	5,921	0.552	0.593	0.727
10	11,364	0.877	0.915	0.884
11	10,232	0.901	0.891	0.727
12	4,466	0.818	0.806	0.746
13	7,130	0.636	0.538	0.500
14	7,320	0.951	0.918	0.812
15	8,532	0.979	0.978	0.936
16	15,928	0.819	0.906	0.635
17	10,012	0.837	0.878	0.900
18	17,956	0.593	0.589	0.321
19	5,960	0.800	0.769	0.588
20	10,874	0.905	0.878	0.794
21	5,425	0.930	0.930	0.948

V. DISCUSSION

A. Overall Model Comparison

At the aggregate level, Claude Sonnet 4.6 and GPT-5.4 are interchangeable for zero-shot CTI NER within measurement precision. The practical choice depends on operational priorities. Claude's higher recall (0.890 vs 0.866) is preferable when missing indicators are more costly than reviewing false positives, which is the typical operational stance for threat intelligence ingestion. GPT-5.4's marginally better macro F1 (0.523 vs 0.514) may reflect better handling of low-frequency entity types such as goal and tool.

LLaMA 4 Scout trails the commercial models at micro F1 = 0.726 zero-shot and 0.632 few-shot. The model shows particular weakness in tool entity precision (0.288) and target sector recall (0.431). As an open-weight model, it remains attractive for data-residency-constrained or air-gapped environments. The performance gap relative to Claude and GPT is substantial but likely

closeable with domain-specific fine-tuning on a CTI NER corpus.

B. Prompt Specificity as a High-Leverage Intervention

The single most impactful change across the study (discussed in the Methodology section), adding hash length constraints, produced larger improvements for the affected entity types than the entire few-shot conditioning effort. This establishes a design principle: for entity types with formally constrained structures (hashes, CVE identifiers, IP addresses), precise format specification in the system prompt is a prerequisite, not a refinement. The same principle motivated the controlled vocabulary intervention for sector and goal, and the explicit CVE format requirement for vulnerability.

However, the MD5 case illustrates the inherent precision/recall trade-off in format-constrained extraction. The revised prompt required strictly lowercase, 32-character hexadecimal strings, which suppressed false positives from non-hash hex sequences but simultaneously caused genuine MD5 values to be missed, subsequently reducing recall.

C. Comparison with Fine-Tuned Baselines

All three models in this study operate without task-specific training, so situating their performance against fine-tuned supervised systems is necessary to interpret the results. No controlled comparison is possible here because the published baselines differ from the present study and from one another, along several axes that each affect the reported score.

TABLE VI. COMPARISON WITH FINE-TUNED CTI NER

Approach	Corpus, taxonomy	Reported F1
This work	21 CTI reports, 15 entities	0.832, 0.831, 0.726 (micro)
[8]	4 harmonized corpora, 21 entities	0.723 to 0.736 (micro)
[2]	DNRTI dataset, 13 entities	0.875 to 0.883 (macro)
[4]	DNRTI dataset, 13 entities	0.900
[5]	7 entity types	0.811
[6]	5 entity types	0.945
[3]	Vulnerability-related entities only	0.968
[7]	16 IoC-related entities	0.975

The single comparison conducted on a comparable footing is with [7], the only baseline evaluated on a broad, cross-source, STIX-aligned taxonomy of 21 entity types against the 15 used in the present study, with entity-level micro F1. On that benchmark, the best fine-tuned transformer-CRF reaches micro F1 0.736, with the remaining encoders clustered between 0.723 and 0.730. The zero-shot micro F1 obtained in the present study, 0.832 for Claude Sonnet 4.6 and 0.831 for GPT-5.4, falls at or above the fine-tuned range on the most similar task, and does so without any task-specific training. The strongest configuration measured, Claude Sonnet 4.6 under few-shot prompting at 0.837, exceeds the fine-tuned range by a clear margin. Even the strongest single-source result reported in [7], an in-domain DNRTI model at 0.871, is achieved on 13 entity types with the model trained and tested on the same corpus and annotation style, a substantially easier setting than the cross-source, fifteen-type extraction evaluated here.

The remaining approaches in Table VI are not directly comparable, and their higher figures do not contradict this finding, because each is produced under conditions more favourable to a fine-tuned model than the present evaluation. References [3] and [5] report 0.875 to 0.883 and 0.900, respectively, but on an in-domain split of the 13-type DNRTI corpus, and the figures in [2] are macro-averaged rather than micro-averaged and so are not directly comparable to the micro F1 reported here. CyBERT reports 0.879 on its own seven-type dataset [5], and SecureBERT 2.0 reaches 0.945 over only five broad categories [6], both on a single in-domain corpus. These results reflect specialization to one annotation style and a narrower taxonomy, not extraction across heterogeneous sources.

The two highest scores in Table VI rest on the least comparable basis and are informative only for a subset of the taxonomy. Reference [3] reports 0.968 on structured vulnerability-description fields such as operating system, vendor, and version rather than CTI-campaign entities, and the models in [7] reach 0.975 over only sixteen indicator-of-compromise types, produced by regular-expression parsers rather than human annotation. Both are relevant to the technical-indicator categories alone, where the present models already perform near-perfectly, and not to overall extraction.

D. Few-Shot Prompting: Hallucination Vs. Cardinality Risk

Few-shot prompting reliably reduces hallucination but introduces cardinality anchoring risk when exemplar entity counts are not representative of the target corpus distribution. The DynoWiper failure demonstrates that this risk is not theoretical – it produces a 73% collapse in micro F1 on an affected report. Practitioners constructing few-shot prompts for CTI NER should 1) sample exemplars to reflect the full entity count distribution per type, 2) include at least one high-cardinality example for any entity type that can appear as a large list, and 3) evaluate the few-shot condition per-report rather than relying solely on corpus-level averages.

Taken together, these findings suggest that the principal bottleneck in prompt-based CTI NER is not model capability in the general sense, but rather the structural properties of the entity types being extracted and the degree to which the prompt communicates their formal constraints. Entities with machine-readable formats (CVE identifiers, IP addresses, cryptographic hashes) are amenable to near-perfect extraction through explicit specification alone. Entities that require operational domain knowledge to disambiguate (tool versus malware, implied sector versus named sector) resist improvement through either few-shot conditioning or format constraints, pointing toward fine-tuning on annotated CTI corpora as the necessary next step. The consistent hierarchy of difficulty across three architecturally distinct models reinforces that this is a domain-level challenge, not a model-level one.

E. Limitations

Several limitations of this study should be acknowledged. The evaluation corpus comprises 21 manually annotated reports. This is a modest scale, and the per-entity-type estimates for low-frequency categories such as goal and tool are therefore subject

to wider variance than the corpus-level figures. The corpus was constructed manually because, to the best of the author's knowledge, no existing public CTI NER dataset covers the 15-type taxonomy evaluated here. Prior corpora target narrower or differently structured schemas, such as the 13 types of DNRTI or the 5 of CyNER, and none of these combine threat-actor attributes, campaign context, and the indicator-of-compromise granularity used in this study within a single annotation scheme. Building a corpus aligned to this taxonomy, therefore, required manual annotation, which bounded its size. A larger corpus spanning more report authors, threat-actor origins, and publication periods would yield more stable per-type estimates and reduce the influence of individual outlier reports on the aggregate metrics, and remains a direction for future work.

All extractions are scored by exact string matching after lowercasing and whitespace normalization, which is a conservative criterion. As documented in Section V, extractions that identify the correct entity with an imprecise boundary, such as "cobalt strike beacon" for "cobalt strike", "atera agent" for "atera", or "fast reverse proxy server" for "fast reverse proxy server (frps)", are counted as both a false positive and a false negative under exact matching, even though they would be operationally useful to an analyst. The reported scores, therefore, represent a lower bound on extraction quality, and a partial-match or soft-F1 metric that credits boundary-imperfect but semantically correct extractions would yield higher figures, particularly for the tool and malware categories where boundary ambiguity is most frequent. Reporting such a metric alongside an exact match is a natural refinement for future evaluation.

A further consideration is potential contamination of the evaluation by data seen during pre-training. Because part of the corpus was published in 2025 and 2026, after the evaluated models' pre-training cutoffs, full memorization is unlikely for those documents, though the underlying campaigns and indicators may still appear in pre-training corpora. This effect is conservative for the baseline comparison in Section V: a prompt-based model may benefit from prior exposure to a report, whereas a fine-tuned encoder trained on a disjoint corpus cannot, so any contamination present would understate rather than inflate the gap. A per-report F1 stratification by publication date is left to future work.

Beyond these immediate constraints, the scope of the evaluation is intentionally limited to named entity identification. Extending the framework to STIX 2.1 relationship extraction and moving beyond the identification of individual entities to the structured representation of relationships between them, such as threat actor-to-malware attribution or campaign-to-infrastructure associations, represents a natural and high-value next step toward fully automated CTI report-to-STIX conversion pipelines.

VI. Conclusion

This study evaluated Claude Sonnet 4.6, GPT-5.4, and LLaMA 4 Scout on CTI named entity recognition across 21 reports, 15 entity types, two prompting conditions, and two prompt versions. Commercial LLMs achieve strong zero-shot CTI NER performance (~0.83 micro F1) without task-specific fine-tuning. A single prompt revision (adding explicit character-length constraints for hash entities) produced the largest F1 improvements in the study, raising SHA-1 F1 by +0.442 (GPT-5.4) and +0.373 (LLaMA), demonstrating the outsized leverage of format specification for technically structured entities.

Few-shot prompting consistently reduces hallucination rates across all models but introduces cardinality anchoring risk, demonstrated most severely by the GPT-5.4's DynoWiper failure. All three models show neutral-to-negative few-shot F1 deltas, with hallucination reduction as the primary measurable benefit of the few-shot condition. The tool entity type is the most challenging across all models due to systematic malware/tool boundary ambiguity, system utility over-extraction, and name normalization failures that exact-match evaluation does not credit.

Collectively, these findings provide both a practical guide for deploying LLM-based CTI NER and an empirical basis for future research in prompt engineering, exemplar selection strategy, and evaluation methodology for the CTI domain.

References

[1] P. Alaeifar, S. Pal, Z. Jadidi, M. Hussain, and E. Foo, "Current approaches and future directions for Cyber Threat Intelligence sharing: A survey," Journal of Information Security and Applications, vol. 83, p. 103786, Jun. 2024, doi: 10.1016/j.jisa.2024.103786.

[2] P. Evangelatos et al., "Named Entity Recognition in Cyber Threat Intelligence Using Transformer-based Models," in 2021 IEEE International Conference on Cyber Security and Resilience (CSR), Rhodes, Greece: IEEE, Jul. 2021, pp. 348–353. doi: 10.1109/CSR51186.2021.9527981.

[3] P. Ranade, A. Piplai, A. Joshi, and T. Finin, "CyBERT: Contextualized Embeddings for the Cybersecurity Domain," in 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA: IEEE, Dec. 2021, pp. 3334–3342. doi: 10.1109/BigData52589.2021.9671824.

[4] S. Zhou, J. Liu, X. Zhong, and W. Zhao, "Named Entity Recognition Using BERT with Whole World Masking in Cybersecurity Domain," in 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), Xiamen, China: IEEE, Mar. 2021, pp. 316–320. doi: 10.1109/ICBDA51983.2021.9403180.

[5] S.-S. Chen, R.-H. Hwang, C.-Y. Sun, Y.-D. Lin, and T.-W. Pai, "Enhancing Cyber Threat Intelligence with Named Entity Recognition Using BERT-CRF," in GLOBECOM 2023 - 2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia: IEEE, Dec. 2023, pp. 7532–7537. doi: 10.1109/GLOBECOM54140.2023.10436853.

[6] E. Aghaei, S. Jain, P. Arun, and A. Sambamoorthy, "Securebert 2.0: Advanced Language Model for Cybersecurity Intelligence," in 2025 Annual Computer Security Applications Conference Workshops (ACSAC Workshops), Honolulu, HI, USA: IEEE, Dec. 2025, pp. 382–388. doi: 10.1109/ACSACW69556.2025.00071.

[7] H.-J. Chan, C.-Y. Hsu, C.-C. Chien, J.-J. Wu, and H.-L. Ku, "FeedRef2022: A Named Entity Recognition Dataset for Extracting Indicators of Compromise," in 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan: IEEE, Dec. 2022, pp. 2578–2584. doi: 10.1109/BigData55660.2022.10020985.

[8] Y. Ech-Chammakhy, A. Motii, A. Rabii, O. Azrara, and J. Chbili, "CyberNER: A Harmonized STIX Corpus for Cybersecurity Named Entity Recognition," in 2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guiyang, China: IEEE, Nov. 2025, pp. 2190–2197. doi: 10.1109/Trustcom66490.2025.00255.

[9] S. Della Penna, R. Natella, V. Orbinato, L. Parracino, and L. Pianese, "CTI-HAL: A Human-Annotated Dataset for Cyber Threat Intelligence Analysis," in 2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Venice, Italy: IEEE, Jun. 2025, pp. 69–78. doi: 10.1109/EuroSPW67616.2025.00014.

[10] K. Guru, R. J. Moss, and M. J. Kochenderfer, "On Technique Identification and Threat-Actor Attribution using LLMs and Embedding Models," in 2025 International Conference on Cybersecurity and AI-Based Systems (Cyber-AI), Varna, Bulgaria: IEEE, Sep. 2025, pp. 332–339. doi: 10.1109/Cyber-AI66431.2025.11233715.

[11] H. Alturkistani, A. G. Jaafar, and S. Chuprat, "Automating Cyber Threat Intelligence Workflows with LLMs: Processing, Analyzing, and Defending in One Model," in 2025 3rd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates: IEEE, Jul. 2025, pp. 1–6. doi: 10.1109/ICCR67387.2025.11292095.

[12] K. Bandla, S. Castro "APTnotes," Source: [Online]. Available: https://github.com/aptnotes/

[13] D. Plohmann, M. Clauß, S. Enders, and E. Padilla, "Malpedia: A Collaborative Effort to Inventorize the Malware Landscape," The Journal on Cybercrime and Digital Investigations, vol. 3, no. 1, pp. 1–19, Dec. 2017, doi: 10.18464/cybin.v3i1.17.

[14] A. Papoutsis et al., "CTI-GEN: A Framework for Generating STIX 2.1 Compliant CTI Using Generative AI," in 2025 IEEE International Conference on Cyber Security and Resilience (CSR), Chania, Crete, Greece: IEEE, Aug. 2025, pp. 334–341. doi: 10.1109/CSR64739.2025.11130126.

A QR-Code-Based Mobile Learning System for Science Instruction in Resource-Limited Schools

Gina B. Selga
College of Education, Pangasinan State University – Asingan Campus, Philippines

***Abstract*—This study developed and evaluated a QR-Code-Based Mobile Learning System for Science Instruction in Resource-Limited Schools. The system was designed to provide junior high school students with low-cost, mobile-accessible science learning materials through printed QR Code Cards linked to a mobile-responsive digital repository. The developed module set consisted of six curriculum-aligned Life Science topics: Cell: The Basic Unit of Life, Cell Theory, Cell Structure and Functions, Prokaryotic vs. Eukaryotic Cells, Cell Types and Cell Modifications, and Cell Cycle. A developmental, quasi-experimental, and descriptive-evaluative research design was employed. The participants were 80 Grade 9 and Grade 10 students divided into an experimental group using the QR-code-based mobile learning system and a control group using traditional printed flashcards. Results showed that students had moderate overall familiarity with QR codes (M = 3.52), highly positive attitudes toward the system (M = 4.45), and high usability ratings (M = 4.38). The QR Code Cards also received very high ratings in terms of portability, durability, readability, and sustainability (M = 4.50). Critically, the experimental group obtained a post-test mean score of 89.15 compared with the control group's mean score of 78.42, yielding a statistically significant difference (t = 6.24, p = .002, Cohen's d = 2.49, 95% CI [7.17, 14.29]), indicating that the QR-code-based mobile learning system significantly improved science achievement by a mean margin of 10.73 points. The findings suggest that QR-code-supported mobile learning can serve as a practical, scalable, and cost-effective instructional innovation for improving science instruction in resource-limited school settings.**

***Keywords*—Academic performance; educational technology; mobile learning; QR code; resource-limited schools; science instruction; usability**

I. Introduction

The increasing demand for technology-supported instruction has encouraged schools to adopt flexible, accessible, and learner-centered approaches that can improve educational delivery across diverse learning environments [1]. In science education, this demand is especially important because students often encounter abstract concepts, microscopic structures, biological processes, and scientific relationships that are difficult to understand through text-based instruction alone [2], [3].

Mobile learning has expanded considerably as a field over the past two decades, with a growing body of evidence documenting its effectiveness in diverse educational contexts [4]. Garzon's review of twenty years of mobile learning in science education found positive effects on student achievement and engagement across multiple grade levels [4]. Traxler argued that mobile learning is particularly transformative when it addresses access inequalities, enabling students without physical libraries or laboratories to engage meaningfully with instructional content [5]. Unlike traditional computer-based learning, mobile learning supports anytime and anywhere access to educational content, enabling students to continue learning beyond the physical boundaries of the classroom [6], [7]. In resource-constrained schools, mobile-responsive systems offer a cost-effective alternative to expensive computer laboratories or printed reference materials. Reychav, Dunaway, and Kobayashi found that students' technology-fit behavior outside the classroom is influenced by the alignment between the mobile platform and actual learning tasks [6]. When mobile learning systems are designed with curriculum alignment and ease of use as central principles, students are more likely to engage with and benefit from these tools.

The effectiveness of mobile learning has been demonstrated in various fields, particularly in health and professional education, where mobile applications have improved students' knowledge, skills, self-efficacy, clinical reasoning, and competency development [8], [9]. Studies on smartphone-based learning applications have also shown that mobile platforms can enhance active learning, collaborative learning, and independent practice when instructional content is designed for meaningful interaction [10], [11].

Mobile-assisted instruction is especially relevant to Generation Z learners, who are accustomed to smartphone use, rapid digital interaction, multimedia content, and technology-mediated communication [9]. However, students' willingness to use digital learning systems depends not only on access to devices but also on perceived usefulness, perceived ease of use, motivation, and technology acceptance [12], [13].

The Technology Acceptance Model explains that learners are more likely to adopt a technology when they believe it is useful and easy to use [13]. Similarly, studies on digital learning behavior suggest that students' engagement with mobile platforms is influenced by technology fit, learning context, and the extent to which the system supports their actual learning needs [6], [12].

Motivation is also a critical factor in the design of mobile learning systems because students are more likely to engage with learning materials when the system captures attention, establishes relevance, builds confidence, and promotes satisfaction [14]. These motivational principles are important in resource-limited schools, where students may experience reduced access to libraries, laboratories, printed references, and other conventional learning resources [1].

In resource-constrained educational settings, mobile learning provides a cost-effective way to extend access to instructional materials without requiring expensive computer laboratories or advanced infrastructure [4], [5]. For schools with limited physical resources, mobile-responsive learning systems can serve as alternative repositories of multimedia content, allowing students to access videos, diagrams, notes, and interactive activities using available smartphones [6], [15].

Quick Response codes, or QR codes, offer a simple and low-cost mechanism for connecting printed materials to digital learning resources [16]. By scanning a QR code with a smartphone camera, learners can immediately access online or mobile-based content without the need for complex software installation or advanced technical skills [17], [18].

QR codes have been widely applied in education because they can transform static printed materials into interactive learning tools [16], [19]. In medical and healthcare education, QR codes have been used to provide students and practitioners with fast access to learning resources, patient education materials, procedural guides, and interactive academic content [17], [18].

A scoping review of QR code use in healthcare education reported that QR codes are useful for delivering accessible, interactive, and context-specific learning materials [19]. Similarly, broader reviews of QR-code-based learning have shown that QR codes can improve student interaction with content, increase perceived learning effectiveness, and support flexible access to educational resources [20].

In basic education, QR-code-enhanced instructional materials have been used to increase learner engagement by combining physical learning prompts with multimedia content [30]. QR-integrated picture cards and science flashcards have shown potential for improving student participation, conceptual understanding, and retention because they allow students to move from static printed information to richer digital explanations [21], [22].

Flashcards remain a valuable instructional tool because they promote retrieval practice, repetition, and long-term memory reinforcement [23], [24]. Retrieval practice strengthens learning by requiring students to recall information actively rather than merely rereading or reviewing content passively [24].

However, traditional flashcards are limited by their static format and may not sufficiently support complex science concepts that require dynamic visualization, detailed explanation, and multiple forms of representation [2], [3]. Cognitive Load Theory suggests that well-designed instructional materials can support learning by reducing unnecessary cognitive demands and providing learners with clearer representations of difficult concepts [3].

The integration of QR codes into flashcards can address these limitations by combining the memory benefits of retrieval practice with the multimedia advantages of mobile learning [21], [22]. Through this approach, a physical card can function as both a review tool and a digital access point to video lectures, three-dimensional illustrations, lecture notes, discussion prompts, and formative activities.

Mobile video-based learning is particularly useful because videos can support independent learning, repeated viewing, and learner control over pace and sequence [15]. When QR codes are used to provide direct access to short instructional videos and visual learning materials, students can engage with science concepts more flexibly and independently [15], [20].

The use of mobile applications and multimedia systems has also been found to improve learning outcomes when students are provided with opportunities for repeated practice, immediate access to information, and interactive engagement [25], [26]. In nursing and health education, mobile applications, augmented reality tools, simulations, and blended mobile learning models have produced positive effects on students' knowledge, skills, motivation, self-efficacy, and competence [27], [28], [29].

Although many of these studies were conducted in higher education or healthcare education, their findings demonstrate the broader instructional value of mobile learning systems for skills development, knowledge acquisition, and learner engagement [8], [30]. These findings suggest that mobile learning principles can also be applied to junior high school science instruction, especially when the system is designed to be simple, accessible, and aligned with curriculum competencies.

Teacher evaluation and student learning outcomes are also important considerations in technology-enhanced instruction because the success of a learning innovation depends on both instructional design and learner response [31], [32]. Survey-based evaluation can provide useful evidence regarding student attitudes, usability, perceived effectiveness, and acceptance of instructional technology when properly designed and reported [33].

In evaluating educational technology interventions, both quantitative and qualitative evidence can strengthen the interpretation of results [7], [8]. Quantitative data can measure learning outcomes, usability, and acceptance, while qualitative data can explain learner experiences, implementation challenges, and perceived value of the system [34], [35].

Thematic analysis is often used to interpret open-ended responses, interview data, and qualitative feedback in educational research [34], [35]. However, researchers must apply systematic procedures to ensure that themes are credible, consistent, and grounded in participant responses [36], [37].

In K-12 science education specifically, mobile learning has shown significant instructional potential. Studies situated in primary and secondary school settings demonstrate that mobile devices can improve science concept acquisition, support inquiry-based learning, and increase student engagement when instructional content is aligned with curriculum competencies [4], [38], [39]. These findings are especially relevant to resource-limited school contexts, where mobile technology may substitute for libraries, laboratories, and other conventional resources that are unavailable or insufficient.

Despite the growing body of research on mobile learning, QR codes, and technology-enhanced education, there remains a need for more studies focused on low-cost mobile learning systems in basic education science classrooms [4], [20]. Much of the available literature focuses on higher education, nursing education, medical education, or general digital learning

behavior, leaving fewer empirical studies situated in resource-limited junior high school contexts [9], [19].

There is also a need to examine how QR-code-based systems can be designed as practical instructional solutions for schools with limited libraries, laboratories, internet infrastructure, and printed learning materials [1], [5]. In such contexts, QR-code-based mobile learning systems may provide an affordable bridge between physical instructional materials and digital learning resources [15], [16].

The present study responds to this research gap by developing and evaluating a QR-code-supported mobile learning system for junior high school science instruction. The system uses printed QR-code cards linked to a mobile-responsive digital repository containing multimedia science materials aligned with curriculum competencies.

Specifically, the study examines students' familiarity with QR-code technology, the usability and physical characteristics of the QR-code cards, students' attitudes toward the mobile learning system, and the system's effect on science academic performance. By combining instructional material design, mobile learning, QR-code integration, technology acceptance, and academic performance evaluation, the study contributes to the growing literature on low-cost educational technology for resource-limited schools.

II. Methodology

A. Research Design

This study employed a developmental and quasi-experimental research design to develop, implement, and evaluate a QR-code-based mobile learning system for junior high school science instruction in a resource-limited school setting. The developmental component focused on the design and production of the QR-code cards and the linked mobile-responsive digital learning repository. The quasi-experimental component examined the effectiveness of the system by comparing the academic performance of students exposed to the QR-code-based mobile learning system with that of students taught using traditional printed flashcards. The study also included a descriptive-evaluative component to assess students' familiarity with QR-code technology, their attitudes toward the system, the usability of the mobile learning materials, and the perceived impact of the physical characteristics of the QR-code cards.

B. Research Locale

The study was conducted at San Vicente West Integrated School, Asingan, Pangasinan, Philippines, a school context characterized by limited instructional resources and the absence of a functional library. This setting was selected because it represented the type of resource-limited learning environment for which the QR-code-based mobile learning system was designed. The lack of sufficient physical learning materials created a need for a low-cost, scalable, and mobile-accessible instructional system that could provide students with supplementary science learning resources.

C. Participants and Sampling Technique

This study was conducted at San Vicente West Integrated School, Asingan, Pangasinan, Philippines. A total population sampling technique was employed, involving all grade 9 and grade 10 students enrolled in life science for the SY. 2025-2026, comprising 40 students from each grade level for a total of 80 participants. Within each grade level, students were randomly assigned into two groups of 20: an experimental group, which used the QR-code-based mobile learning system, and a control group, which used traditional printed flashcards. This resulted in a combined experimental group of 40 students (20 from grade 9 and 20 from grade 10) and a combined control group of 40 students (20 from grade 9 and 20 from grade 10), ensuring that both groups were comparable in terms of grade-level composition before the intervention.

D. The QR-Code-Based Mobile Learning System

The instructional intervention developed for this study was a QR-code-based mobile learning system composed of two major components: printed QR-code cards and a mobile-responsive digital learning repository. Each QR-code card was produced using 3 × 5-inch laminated cardstock with embedded QR codes linked to specific science learning modules. The system used static QR codes—fixed codes linked to stable, persistent URLs hosted on the digital repository server. Because static QR codes encode the URL directly within the code itself, the physical cards must be reprinted if the linked URL changes. To mitigate this limitation, the researcher ensured that the repository domain and URL structure remained unchanged throughout the study period and documented a link-preservation protocol for future use. Future implementations may consider dynamic QR codes, which allow URL redirection at the database level without requiring the reprinting of physical cards. The six science modules covered were: Cell: The Basic Unit of Life, Cell Theory, Cell Structure and Functions, Prokaryotic vs. Eukaryotic Cells, Cell Types and Cell Modifications, and Cell Cycle.

E. Development Procedure

The development of the QR-code-based mobile learning system followed four phases: needs analysis, content development, system design, and validation. During the needs analysis phase, the researcher identified science topics that students commonly found difficult to understand using printed materials alone. Priority was given to abstract biological concepts that required visual explanation, such as cell structure, cell functions, and the cell cycle. During the content development phase, the researcher prepared and curated multimedia resources aligned with the DepEd MELCs. Each module was designed to include concise explanations, visual materials, and activities that could support independent and classroom-based learning. During the system design phase, QR codes were generated and embedded into the printed cards. Each QR code was linked to a corresponding page in the mobile-responsive digital repository. The physical cards were designed with high-contrast text, visual cues, and module-specific illustrations to support readability, scanning accuracy, and learner engagement. During the validation phase, the QR-code cards, digital learning materials, and research instruments were evaluated by three validators: a content expert, a technical or

ICT expert, and a language and research expert. The validators assessed the materials in terms of content accuracy, technical quality, layout and design, usability, and clarity. Revisions were made based on the validators' comments before the system was implemented in the classroom.

F. Research Instrument

The study used the following instruments:

1) Science achievement test. A researcher-developed pre-test and post-test were used to measure students' science achievement before and after the intervention. The test covered the science topics included in the QR-code-based mobile learning system.

2) QR-Code technology familiarity survey. This instrument measured students' prior exposure to QR codes in daily life, social media, and academic contexts.

3) Usability and attitude survey. This questionnaire assessed students' perceptions of the QR-code-based mobile learning system in terms of ease of use, usefulness, accessibility, engagement, and willingness to use the system for science learning.

4) Physical characteristics evaluation form. This instrument measured students' evaluation of the QR-code cards in terms of size, color, texture, durability, portability, readability, and sustainability.

5) Development difficulty rating form. This instrument assessed the level of difficulty encountered in developing the QR-code cards and the mobile-responsive digital repository.

All survey instruments used a five-point Likert scale. The instruments were reviewed by experts and pilot-tested before implementation to ensure clarity, relevance, and appropriateness for the target respondents. Fig. 1 shows the system architecture of the QR-Code-Based Mobile Learning System.

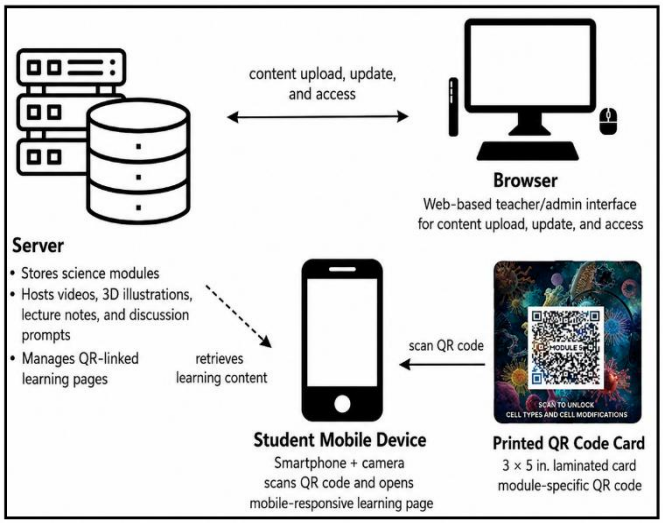


Fig. 1. Architecture of the QR-code-based mobile learning system.

Fig. 1 illustrates the system architecture of the QR-code-based mobile learning system developed for science instruction in a resource-limited school setting. The system architecture is composed of a server module, a browser-based interface, the student's mobile device, and the printed QR Code Card. These components work together to connect physical instructional materials with digital science learning resources.

The Server Module serves as the core storage and management component of the system. It houses the digital learning repository containing the science instructional materials developed for the six modules included in the study, namely: Cell: The Basic Unit of Life, Cell Theory, Cell Structure and Functions, Prokaryotic vs. Eukaryotic Cells, Cell Types and Cell Modifications, and Cell Cycle. The repository stores the multimedia learning resources linked to each QR code, such as short video lectures, three-dimensional biological illustrations, lecture notes, and interactive discussion prompts. In addition to storing content, the server module manages the links between the QR codes and their corresponding mobile learning pages. This makes the system function as a centralized digital platform that supports the delivery of science content through mobile access.

The Browser represents the web-based interface used to access and manage the digital repository. In the context of the study, this browser component may be used by the teacher or researcher to upload, organize, and update the content for each science module. Because the learning repository is mobile-responsive, the browser-based system allows the instructional content to be delivered in a format that is accessible on standard web-enabled devices without the need for a dedicated application. This component is important because it supports the maintainability and scalability of the system. Once a module page is updated in the repository, the linked QR code can continue to direct students to the revised content without requiring the printed card to be replaced, especially if dynamic QR linking is used.

The Mobile Device refers to the student's smartphone, which functions as the primary access tool for the mobile learning system. Using the phone's built-in camera, the student scans the QR code printed on the card. After scanning, the smartphone opens the linked learning page where the student can view digital science resources and participate in related learning activities. This design is especially appropriate for resource-limited schools because it makes use of widely available student smartphones rather than requiring computer laboratories or specialized learning hardware. The mobile device, therefore, acts as the bridge between the printed card and the digital repository, enabling the system to support both classroom learning and independent study outside school hours.

The QR Code Card is the physical instructional component of the system. Based on the manuscript, each card was produced using 3 × 5-inch laminated 220 gsm cardstock and designed for portability, durability, readability, and repeated classroom use. Each card contains a module-specific QR code that directs students to the corresponding science content in the digital repository. The QR Code Card functions not only as a printed learning aid but also as a physical-digital access point. In this sense, it transforms a conventional flashcard into an interactive learning object by integrating retrieval-based review with mobile multimedia learning.

G. Data Gathering Procedure

Before data collection, administrative approval was secured from the school principal, and informed consent was obtained from the students and their guardians. The participants were informed about the purpose of the study, the procedures involved, and the confidentiality of their responses. The data gathering procedure began with the administration of the pre-test to both the experimental and control groups. The pre-test established the students' baseline level of science achievement before the instructional intervention. After the pre-test, the experimental group was oriented on how to use the QR-code-based mobile learning system. Students were taught how to scan the QR codes, access the mobile learning repository, navigate the multimedia resources, and use the cards during classroom and independent study activities. The experimental group then used the QR-code-based mobile learning system during science instruction. The students accessed video lectures, biological illustrations, notes, and discussion prompts through the QR-code cards. The control group, on the other hand, received instruction using traditional printed flashcards covering the same science topics. After the intervention period, both groups took the post-test. The experimental group also answered the usability, attitude, QR-code familiarity, and physical characteristics evaluation questionnaires. The collected data were then encoded, tabulated, and analyzed using appropriate statistical tools.

H. Data Analysis

Descriptive statistics were used to summarize the respondents' profiles, QR-code familiarity, usability ratings, attitude toward the system, physical card evaluation, and perceived development difficulty. Frequency counts, percentages, means, standard deviations, and weighted means were computed. The weighted means were interpreted using a five-point verbal interpretation scale. Higher mean scores indicated more positive perceptions of the QR-code-based mobile learning system, greater usability, stronger technology familiarity, or higher perceived impact of the physical characteristics of the cards. To establish baseline equivalence before the intervention, an independent-samples t-test was conducted on the pre-test scores of the experimental and control groups. To determine effectiveness, the post-test scores were compared using an independent-samples t-test. Effect size was calculated using Cohen's d and a 95% confidence interval was computed for the mean difference in post-test scores, enabling practical significance assessment alongside statistical significance. This test determined whether there was a statistically significant difference in science achievement between students who used the QR-code-based mobile learning system and those who used traditional printed flashcards.

I. Proposed Methodological Framework

Fig. 2 presents the methodological framework of the study, organized into three sequential phases. It shows that the research follows three major phases: system development, implementation, data gathering, and data analysis and interpretation.

In Phase 1, the QR-code-based mobile learning system is developed by identifying difficult science topics, preparing multimedia learning content, designing the mobile-responsive repository and QR code cards, and validating the materials through experts. In Phase 2, the system is implemented with junior high school students. The participants are divided into an experimental group using the QR-code-based mobile learning system and a control group using traditional printed flashcards. Pre-tests, post-tests, and survey instruments are administered to collect data on learning performance, usability, attitude, QR familiarity, and card characteristics. In Phase 3, the collected data are analyzed using descriptive statistics and an independent-samples t-test. The findings are then interpreted to determine whether the QR-code-based mobile learning system is usable, acceptable, and effective in improving science instruction in resource-limited schools.

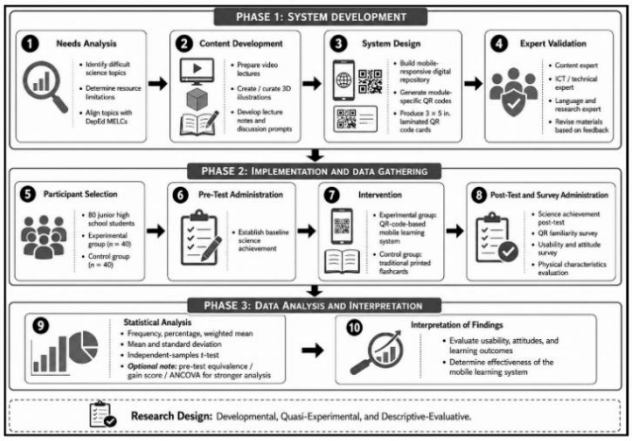


Fig. 2. The proposed methodological framework.

J. Ethical Consideration

The study followed ethical procedures for school-based research. Administrative clearance was obtained before data collection, and informed consent was secured from the participants and their guardians. Participation was voluntary, and students were informed that their responses would be used only for research purposes. No personally identifiable information was included in the analysis or reporting of results. The study also ensured that both the experimental and control groups received science instruction aligned with the same curriculum competencies. The QR-code-based learning materials were used for educational purposes only.

III. Results and Discussion

A. Profile of the Respondents

The respondents of the study were junior high school students from San Vicente West Integrated School, Asingan, Pangasinan, Philippines. The sample consisted of Grade 9 and Grade 10 students, with a total of 80 participants divided into two groups: the experimental group using the QR-code-based mobile learning system and the control group using traditional printed flashcards. Each group had 40 students. Most of the respondents were between 14 and 16 years old, representing 88% of the sample. The sex distribution was nearly balanced, with 52% female students and 48% male students. Before the intervention, 65% of the students were within the "Satisfactory" range in science, with grades between 75% and 79%. This profile indicates that the participants were appropriate for the intervention because they represented learners who could

benefit from additional instructional support in science. The baseline profile also supports the need for a mobile learning intervention. Since the school setting was resource-limited and lacked a functional library, students required an accessible instructional system that could supplement classroom learning without depending on expensive infrastructure. The use of smartphones and QR-code cards, therefore, provided a practical way to deliver science learning materials through available technology.

B. Pre-Test Equivalence of the Experimental and Control Groups

Before implementing the intervention, an independent-samples t-test was conducted on the pre-test scores of the experimental and control groups to verify that the two groups were academically equivalent at baseline. Table I presents the pre-test equivalence results.

The pre-test results showed no statistically significant difference between the experimental group ($M = 76.45$, $SD = 4.82$) and the control group ($M = 75.88$, $SD = 5.10$), $t(78) = 0.41$, $p = .684$. The failure to reject the null hypothesis confirms that the two groups were academically equivalent in science before the intervention, thereby strengthening the internal validity of the quasi-experimental design and supporting the interpretability of the post-test comparison.

TABLE I. PRE-TEST EQUIVALENCE OF EXPERIMENTAL AND CONTROL GROUPS

Group	n	Pre-test Mean	SD	t-val	p-val	Decision
Experimental Group: QR-Code-Based Mobile Learning System	40	76.45	4.82	0.41	0.684	Fail to Reject H_0
Control Group: Traditional Printed Flashcards	40	75.88	5.1			

C. Developed QR Code Card Module Set

One of the major outputs of the study was the development of a six-module QR Code Card set for junior high school science instruction. The module set was designed as the physical-digital component of the QR-code-based mobile learning system. Each card contained a QR code that directed students to a mobile-responsive digital repository containing video lectures, three-dimensional biological illustrations, lecture notes, and interactive discussion prompts.

The six QR Code Cards were produced using 3 × 5-inch laminated cardstock with a 220 gsm material specification. The cards were designed to be portable, durable, reusable, and suitable for repeated classroom use. The total production cost of the complete six-module set was reported at ₱10,000.00, covering cardstock, printing, lamination, and hosting. Fig. 3 presents the actual QR Code Card module set developed for the study. The cards were designed with science-themed visual backgrounds, embedded QR codes, and module labels to help students identify the topic before scanning. The QR codes served as access points to the digital learning repository, allowing the printed cards to function as interactive learning materials rather than ordinary flashcards.

This QR Code Card of Fig. 3 introduces the cell as the basic structural and functional unit of life. This module provides students with foundational knowledge about why all living organisms are composed of cells and how cells perform essential life processes. Through the QR-linked multimedia materials, students can access visual explanations, lecture notes, and supporting activities that help them understand the cell as the foundation of biological organization.

This QR Code Card of Fig. 4 focuses on the Cell Theory and its historical development. This module guides students in understanding the major postulates of Cell Theory, including the idea that all living things are made of cells, that the cell is the basic unit of life, and that all cells come from pre-existing cells. The QR-linked resources help students connect scientific discoveries with the development of modern biological understanding.

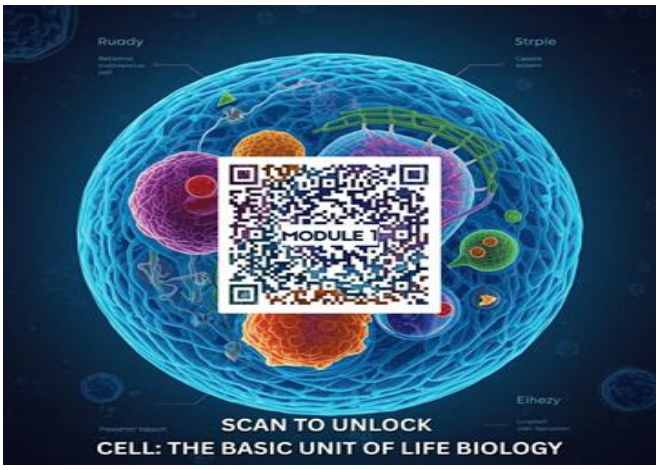


Fig. 3. The cell: The basic unit of life QR code card.

Fig. 4. The cell theory QR code card.

This QR Code Card of Fig. 5 presents the different parts of the cell and their corresponding functions. This module helps students identify major cell structures such as the nucleus, cytoplasm, cell membrane, mitochondria, and other organelles.

Fig. 5. The cell structure and functions QR code card.

The QR-linked videos and three-dimensional illustrations are especially useful in this module because they allow students to visualize cell parts that are normally difficult to observe in a resource-limited classroom.

Fig. 6. The prokaryotic vs. eukaryotic cells QR code card.

This QR Code Card of Fig. 6 helps students distinguish between prokaryotic and eukaryotic cells. This module emphasizes the differences in cell organization, presence or absence of a nucleus, complexity of internal structures, and examples of organisms under each cell type.

Through the QR-linked digital materials, students can compare the two cell types more clearly using diagrams, explanations, and guided activities.

This QR Code Card of Fig. 7 focuses on cell types and cell modifications. This module helps students classify different types of cells and understand how specialized structures support specific cell functions. It is particularly useful for explaining how cells become adapted for particular roles in living organisms. The QR-linked content supports deeper understanding by providing visual examples of specialized cells and their modified structures.

This QR Code Card of Fig. 8 discusses the stages of the cell cycle and their importance in growth, repair, and reproduction. This module introduces students to the sequence of events involved in cell division, including interphase and the major stages of mitosis. Since the cell cycle is a process-based topic, the QR-linked videos and visual materials help students follow the sequence of changes more easily than s printed explanations alone.

Fig. 7. The cell types and cell modifications QR code card.

Fig. 8. The cell cycle QR code card.

Together, the six QR Code Cards form a structured Life Science learning sequence. The modules begin with the basic concept of the cell, proceed to cell theory and cell structures, compare major cell types, examine cell specialization, and conclude with the cell cycle. This progression supports gradual concept development and allows students to access multimedia learning materials for each topic through a simple smartphone scan.

D. QR Code Card Module Set and Curriculum Alignment

Table II presents the six QR code card module sets and their alignment to the curriculum. The six-module set shows that the QR-code-based mobile learning system was not developed as a general or unstructured digital tool. Instead, it was organized around specific curriculum competencies in junior high school Life Science. This alignment is important because it ensures that the mobile learning system supports required learning outcomes rather than functioning only as a supplementary technology activity.

TABLE II. SIX QR CODE CARD MODULE SET AND CURRICULUM ALIGNMENT

Module	Topic	DepEd MELC Competency	Grade Level
1	Cell: The Basic Unit of Life	Describe the cell as the basic structural and functional unit of life	Grade 9 – Life Science
2	Cell Theory	Explain the postulates of the Cell Theory and their historical development	Grade 9 – Life Science
3	Cell Structure and Functions	Identify the parts of the cell and their corresponding functions	Grade 9 – Life Science
4	Prokaryotic vs. Eukaryotic Cells	Distinguish between prokaryotic and eukaryotic cells	Grade 9 – Life Science
5	Cell Types and Cell Modifications	Classify cells and describe how specialized structures support specific cell functions	Grade 10 – Life Science
6	Cell Cycle	Explain the stages of the cell cycle and their significance to growth and reproduction	Grade 10 – Life Science

The module topics also reflect science concepts that often require visual explanation. Topics such as cell structure, cell functions, prokaryotic and eukaryotic cells, and the cell cycle are difficult to teach using text alone because students need to visualize microscopic structures and bsiological processes. The integration of QR-linked videos, diagrams, notes, and discussion tasks allowed each printed card to become an access point to richer science learning resources. The physical design of the module set also contributed to its suitability for resource-limited schools. Since the cards were laminated and reusable, the system could be used across multiple lessons and learning activities. The use of dynamic QR codes also allowed the linked content to be updated without reprinting the cards, increasing the maintainability of the instructional system.

E. Students' Familiarity with QR Code Technology

Table III shows the students' familiarity with QR code technology. The students reported high familiarity with QR codes in daily life and social media contexts, but low familiarity with QR codes as part of formal academic learning.

TABLE III. EXTENT OF STUDENT FAMILIARITY WITH QR CODE TECHNOLOGY

Indicator	Weighted Mean	Interpretation
Actual encounter with QR codes in daily life	4.12	High Familiarity
Use of QR codes in social media platforms	4.35	High Familiarity
Awareness of QR codes within the school curriculum	2.1	Low Familiarity
Overall Mean	3.52	Moderate Familiarity

The results indicate that students were already familiar with QR codes as part of everyday digital practices, particularly in social media and commercial transactions. However, their low familiarity with QR codes in the school curriculum shows that the use of QR codes as an instructional tool was relatively new to them. This finding confirms the need for orientation before implementation. Although the learners were familiar with the scanning process, they had limited experience using QR codes for structured academic learning. This result supports the relevance of the QR-code-based mobile learning system. The system builds on students' existing digital habits and redirects them toward educational purposes. In this way, the intervention transforms a familiar technology into a purposeful learning tool. The moderate overall familiarity also suggests that students were not technologically unprepared, making the system feasible for classroom implementation.

F. Evaluation of the Physical Characteristics of the QR Code Cards

The QR Code Cards were evaluated in terms of size, color, texture, and sustainability. These characteristics were important because the physical card served as the primary access point to the mobile learning system. Table IV shows the impact ratings of QR code card physical characteristics.

TABLE IV. IMPACT RATINGS OF QR CODE CARD PHYSICAL CHARACTERISTICS

Physical Characteristic	Weighted Mean	Interpretation
Size – 3 × 5-inch portability and handling	4.56	Very High Impact
Color – high-contrast design for readability	4.2	High Impact
Texture – laminated durability	4.45	High Impact
Sustainability – re-usability over the full academic year	4.78	Very High Impact
Overall Mean	4.5	Very High Impact

The overall mean of 4.50 indicates that the students evaluated the physical characteristics of the QR Code Cards as having a very high impact. Among the indicators, sustainability received the highest mean of 4.78, followed by size and portability with a mean of 4.56. These results show that the students found the cards durable, reusable, and convenient to handle. The high rating for sustainability is particularly important in a resource-limited school setting. Since the cards were laminated and produced using 220 gsm cardstock, they were able to withstand repeated use during classroom activities. This makes the system practical for schools with limited funds because the cards can be reused over time instead of being replaced frequently. The high rating for size also suggests that the 3 × 5-inch format was appropriate for student use. The card was small enough to be portable but large enough to display the QR code and module information clearly. This supports the idea that the physical design of the learning tool affects its usability and acceptance.

G. Level of Difficulty in Developing the QR-Code-Based Mobile Learning System

The development of the QR-code-based mobile learning system was rated as moderately difficult, with an overall mean of 3.10. The main challenges involved creating a mobile-responsive digital repository, ensuring that the QR codes were scannable across different smartphone models, and preparing

multimedia content aligned with the DepEd Most Essential Learning Competencies. This finding shows that the system was feasible to develop, although it required technical and instructional preparation. The moderate difficulty rating suggests that teachers may be able to replicate the system if they are given proper training in QR code generation, mobile content organization, basic website management, and multimedia curation. The result also highlights the importance of teacher capacity-building. For resource-limited schools to adopt this kind of mobile learning system successfully, teachers need support not only in content development but also in the technical aspects of designing QR-linked learning materials. The inclusion of a teacher's manual or implementation guide may therefore improve the scalability of the system.

H. Students' Attitude Toward and Utilization of the QR-Code-Based Mobile Learning System

The experimental group reported a highly positive attitude toward the QR-code-based mobile learning system, with a mean of 4.45. Students also gave high usability ratings, with a mean of 4.38. These results indicate that the students found the system useful, accessible, and easy to use. Students appreciated the immediacy of access to multimedia learning materials. By scanning the QR Code Cards, they could access video lectures, biological illustrations, lecture notes, and discussion prompts through their smartphones. This made the learning experience more interactive compared with traditional printed flashcards. The system also functioned as a portable learning resource. Since the students could bring the QR Code Cards home, they were able to review the science modules outside the classroom. This is especially valuable in a school without a functional library because the QR-code-based system provided students with access to supplementary science resources beyond regular classroom instruction. The positive attitude and high usability rating suggest that the system was acceptable to the learners. From a technology acceptance perspective, this means that the students were more likely to use the system because they perceived it as easy to scan, easy to navigate, and helpful for learning science concepts.

I. Comparative Academic Performance of the Experimental and Control Groups

To determine the effectiveness of the QR-code-based mobile learning system, the post-test scores of the experimental and control groups were compared using an independent-samples t-test. Table V shows the independent-samples t-test on post-test academic performance.

TABLE V. INDEPENDENT SAMPLE T-TEST RESULT FOR POST-TEST ACADEMIC PERFORMANCE

Group	n	Post-test Mean	SD	t-val	p-val	Decision
Experimental Group: QR-Code-Based Mobile Learning System	40	89.15	3.45	6.24	0.002	Reject H_0
Control Group: Traditional Printed Flashcards	40	78.42	5.12			

The results show that the experimental group obtained a higher post-test mean score of 89.15 compared with the control group's mean score of 78.42. The computed t-value of 6.24 and p-value of .002 indicate a statistically significant difference between the two groups. Therefore, the null hypothesis of no significant difference was rejected. This result suggests that students who used the QR-code-based mobile learning system performed significantly better in science than those who used traditional printed flashcards. The mean difference of 10.73 percentage points indicates a substantial improvement in favor of the experimental group. The higher performance of the experimental group may be attributed to the integration of multimedia resources into the learning process. Unlike traditional flashcards, the QR Code Cards provided access to videos, three-dimensional biological illustrations, notes, and discussion activities. These resources likely helped students visualize abstract science concepts more clearly and review the lessons more independently. The result also supports the instructional value of combining physical learning materials with mobile technology. The printed QR Code Card served as a simple and familiar object, while the linked digital repository expanded the learning experience through multimedia content. This combination made the system suitable for a resource-limited school because it did not require expensive technology infrastructure.

J. Discussions

The findings of the study demonstrate that the QR-code-based mobile learning system is a feasible, usable, and effective instructional innovation for science education in a resource-limited school. The system addressed a specific instructional problem: the lack of sufficient learning resources in a school without a functional library. By connecting printed QR Code Cards to a digital learning repository, the system provided students with access to curriculum-aligned multimedia materials.

The six QR Code Card modules further strengthen the system's instructional value because they were organized around specific Life Science competencies for Grades 9 and 10. The modules covered cell biology topics that require visual and conceptual support, making them appropriate for QR-linked multimedia instruction. Through this structure, the system served not only as a technology innovation but also as a curriculum-aligned instructional intervention.

The pre-test findings established that the experimental and control groups were academically comparable before the implementation of the intervention. Results revealed no statistically significant difference between the experimental group and the control group. This result strengthened the internal validity of the quasi-experimental design because any differences observed in the post-test outcomes could be more confidently attributed to the QR-code-based mobile learning intervention rather than pre-existing academic disparities between the groups.

The students' moderate familiarity with QR code technology shows that the system was appropriate for their digital readiness. Although students were already familiar with QR codes in social media and daily life, they had limited exposure to QR codes as

academic tools. The intervention, therefore, introduced a familiar technology in a new educational context.

The high evaluation of the physical card characteristics confirms that usability is not limited to the digital component of the system. The durability, portability, readability, and reusability of the printed cards contributed to the successful implementation of the intervention. This is particularly important for schools with limited resources because instructional materials must be affordable and long-lasting.

The positive attitude and high usability ratings further indicate that students accepted the QR-code-based mobile learning system. Their favorable response suggests that mobile learning tools are more likely to succeed when they are easy to access, simple to use, and directly connected to classroom learning needs.

Most importantly, the significant difference in post-test performance provides evidence that the system contributed to improved science achievement. The QR-code-based mobile learning system allowed students to access multiple representations of science concepts, including visual, textual, and video-based explanations. These features likely reduced the difficulty of learning abstract topics and supported better understanding.

K. Implication of the Findings

The findings have several implications for science instruction and educational technology implementation. First, QR-code-based mobile learning systems can be used as a practical alternative to expensive digital learning platforms. Since the system only requires printed QR Code Cards, a mobile-responsive repository, and smartphones, it is feasible for schools with limited infrastructure. Second, the system can help compensate for the lack of physical learning resources. In the absence of a functional library, QR Code Cards can serve as portable access points to digital learning materials. Third, the six-module QR Code Card set demonstrates that mobile learning systems should be closely aligned with curriculum competencies. The value of the system does not come from QR codes alone, but from the quality, relevance, and organization of the learning materials connected to each code. Fourth, the study highlights the importance of combining instructional design with technology design. The success of the system depended on the quality of the multimedia content, the alignment with curriculum competencies, the usability of the digital repository, and the physical durability of the printed QR Code Cards. Fifth, the results suggest that teacher training is necessary for wider implementation. Teachers need skills in content curation, QR code generation, mobile-responsive content organization, and evaluation of digital learning materials. Finally, the system may be replicated in other subject areas such as Mathematics, English, and Technology and Livelihood Education. Its low-cost and reusable format makes it suitable for broader school-based adoption.

IV. Conclusion and Recommendations

A. Conclusion

This study concluded that a QR-Code-Based Mobile Learning System for Science Instruction in Resource-Limited Schools is a practical, acceptable, and effective instructional innovation for junior high school science education. The developed system, composed of six QR Code Card modules linked to mobile-responsive multimedia resources, provided students with accessible learning materials despite the limited physical resources of the school. The findings showed that students were moderately familiar with QR code technology, highly receptive to the use of the system, and positive in their evaluation of the QR Code Cards' usability, portability, durability, and sustainability. The significant difference in post-test performance between the experimental and control groups further indicates that the QR-code-based mobile learning system contributed to improved science achievement. Overall, the system demonstrates how low-cost mobile technology can support science instruction, independent learning, and curriculum-aligned resource delivery in resource-limited school settings.

B. Recommendations

Based on the findings, it is recommended that schools with limited instructional resources adopt the QR-code-based mobile learning system as a supplementary tool for science instruction. The system may help address the lack of physical libraries, laboratories, and printed learning materials by providing students with direct access to digital resources through QR Code Cards. Since the cards are low-cost, portable, reusable, and easy to use, they can serve as practical learning materials for classroom activities, independent study, and home-based review.

It is also recommended that the system be expanded to other science topics and subject areas. Future versions may include additional modules in biology, chemistry, physics, mathematics, English, and other curriculum areas where students need visual, interactive, and multimedia-supported learning materials. Expanding the system will help determine whether the QR-code-based mobile learning approach is effective across different subjects and grade levels.

Teacher training should also be provided to support the effective implementation of the system. Teachers need basic skills in QR code generation, mobile-responsive content preparation, multimedia curation, and integration of QR-linked activities into classroom instruction. With proper training, teachers can develop their own QR Code Cards and adapt the system to the learning needs of their students.

Schools are also encouraged to improve digital access for students. Since the system depends on mobile devices and internet connectivity, administrators may consider providing offline digital repositories, local servers, shared internet access, or downloadable learning materials for students with limited mobile data or unstable internet connections.

These measures will help ensure that all learners can benefit from the mobile learning system regardless of their connectivity situation.

Further research is recommended using a wider sample and a more rigorous research design. Future studies may involve multiple schools, larger groups of learners, and randomized or matched-group designs to improve the generalizability of the findings. Researchers may also include long-term learning

measures such as delayed post-tests, QR scan frequency, student engagement analytics, and retention of science concepts over time.

Finally, the mobile learning repository may be enhanced by adding assessment and feedback features. Future versions of the system may include quizzes, progress tracking, automatic feedback, and teacher monitoring tools. These improvements can make the QR-code-based mobile learning system more interactive, data-driven, and useful for both students and teachers.

ACKNOWLEDGMENT

The researcher gratefully acknowledges Pangasinan State University for its support in the conduct of this research. Artificial intelligence tools were used to assist with language refinement, idea organization, and formatting. However, all research processes, findings, interpretations, and conclusions were carefully reviewed and verified by the researcher to ensure accuracy, originality, and academic integrity.

REFERENCES

[1] Department of Education, The Digital Rise Program: Strengthening ICT-Assisted Teaching and Learning in the K-12 Curriculum, DepEd Order No. 042, s. 2022. Republic of the Philippines, 2022.

[2] R. Lazarowitz, "Biology teaching and learning in secondary schools: Students' attitudes and interests," Journal of Biological Education, vol. 44, no. 2, pp. 54–62, 2010, doi: 10.1080/00219266.2010.9656196.

[3] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cognitive Science, vol. 12, no. 2, pp. 257–285, 1988, doi: 10.1207/s15516709cog1202_4.

[4] J. Garzon, "An overview of twenty years of mobile learning in science education," Journal of Educational Computing Research, vol. 59, no. 4, pp. 641–669, 2021, doi: 10.1177/0735633120972022.

[5] J. Traxler, "Learning in a mobile age," International Journal of Mobile and Blended Learning, vol. 1, no. 1, pp. 1–12, 2009, doi: 10.4018/jmbl.2009010101.

[6] I. Reychav, M. Dunaway, and M. Kobayashi, "Understanding mobile technology-fit behaviors outside the classroom," Computers & Education, vol. 87, pp. 142–150, 2015, doi: 10.1016/j.compedu.2015.04.005.

[7] C. Strandell-Laine, H. Leino-Kilpi, E. Löyttyniemi, L. Salminen, M. Stolt, R. Suomi, and M. Saarikoski, "A process evaluation of a mobile cooperation intervention: A mixed methods study," Nurse Education Today, vol. 80, pp. 1–8, 2019, doi: 10.1016/j.nedt.2019.05.037.

[8] S. Jang and E. E. Suh, "Development and application of a mobile-based multimedia nursing competency evaluation system for nursing students: A mixed-method randomized controlled study," Nurse Education in Practice, vol. 64, Art. no. 103458, 2022, doi: 10.1016/j.nepr.2022.103458.

[9] J. H. Kim and H. Park, "Effects of smartphone-based mobile learning in nursing education: A systematic review and meta-analysis," Asian Nursing Research, vol. 13, no. 1, pp. 20–29, 2019, doi: 10.1016/j.anr.2019.01.005.

[10] S.-J. Kim, H. Shin, J. Lee, S. Kang, and R. Bartlett, "A smartphone application to educate undergraduate nursing students about providing care for infant airway obstruction," Nurse Education Today, vol. 48, pp. 145–152, 2017, doi: 10.1016/j.nedt.2016.10.006.

[11] H. Lee, C. Kim, H. Kim, J. Kim, and M. Kang, "Improving active collaborative clinical learning through a mobile application for undergraduate nursing students," Computers, Informatics, Nursing, vol. 10, pp. 1097, 2022, doi: 10.1097/CIN.0000000000000918.

[12] S. F. Persada, B. A. Miraja, and R. Nadlifatin, "Understanding the Generation Z behavior on D-learning: A unified theory of acceptance and use of technology (UTAUT) approach," International Journal of Emerging Technologies in Learning, vol. 14, no. 5, 2019, doi: 10.3991/ijet.v14i05.9993.

[13] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

[14] J. M. Keller, "Motivational design research and development," in Motivational Design for Learning and Performance: The ARCS Model Approach. Boston, MA, USA: Springer, 2009, pp. 297–323.

[15] Z. Pi, J. Hong, J. Yang, and Y. Ning, "Mobile video-based learning: How QR codes facilitate independent learning in vocational settings," Frontiers in Education, vol. 10, Art. no. 114521, 2025, doi: 10.3389/feduc.2025.114521.

[16] H. Crompton, J. LaFrance, and M. van't Hooft, "QR codes 101," Learning & Leading with Technology, vol. 39, pp. 22–25, 2011.

[17] J. Bellot, K. Shaffer, and M. Wang, "Use of quick response coding to create interactive patient and provider resources," Journal of Nursing Education, vol. 54, pp. 224–227, 2015, doi: 10.3928/01484834-20150318-07.

[18] K. Brodie, L. L. Madden, and C. A. Rosen, "Applications of quick response (QR) codes in medical education," Journal of Graduate Medical Education, vol. 12, pp. 138–140, 2020, doi: 10.4300/JGME-D-19-00516.1.

[19] C. T. Karia, A. Hughes, and S. Carr, "Uses of quick response codes in healthcare education: A scoping review," BMC Medical Education, vol. 19, pp. 1–14, 2019, doi: 10.1186/s12909-019-1876-4.

[20] V. Hernandez-Beltran, A. Manzano-Leon, C. Sanchez-Cruzado, and M. A. Sanchez-Compaña, "QR codes as a learning tool in the classroom: A systematic review," International Journal of Educational Technology in Higher Education, vol. 20, no. 1, pp. 1–29, 2023, doi: 10.1186/s41239-023-00392-8.

[21] R. Axelino and S. Ahmad, "The development of interactive picture cards based on QR code for elementary school students," Journal of Physics: Conference Series, vol. 1788, no. 1, Art. no. 012050, 2021, doi: 10.1088/1742-6596/1788/1/012050.

[22] P. Sari, N. Hidayah, and R. Wulandari, "Integration of QR codes in science flashcards: Enhancing student engagement and retention," Asian Journal of Science Education, vol. 5, no. 2, pp. 120–135, 2023.

[23] H. Ebbinghaus, Memory: A Contribution to Experimental Psychology, H. A. Ruger and C. E. Bussenius, Trans. New York, NY, USA: Teachers College Press, 1885.

[24] J. D. Karpicke and J. R. Blunt, "Retrieval practice produces more learning than elaborative studying with concept mapping," Science, vol. 331, no. 6018, pp. 772–775, 2011, doi: 10.1126/science.1199327.

[25] H. Kim and E. E. Suh, "The effects of an interactive nursing skills mobile application on nursing students' knowledge, self-efficacy and skills performance: A randomized controlled trial," Asian Nursing Research, vol. 12, no. 1, pp. 17–25, 2018, doi: 10.1016/j.anr.2018.01.001.

[26] A. Li, N. M. Zain, A. W. Yousef, A. Yusuf, and H. Deng, "Effectiveness of online and offline blended teaching methods using a mobile application on surgical nursing apprenticeship: A clustered randomised controlled trial," Nurse Education in Practice, Art. no. 104320, 2025, doi: 10.1016/j.nepr.2025.104320.

[27] S. Özdemir and Ö. Özfırat, "Vaginal breech birth management: Serious mobile game design and evaluation for midwifery students," Nurse Education Today, vol. 146, Art. no. 106563, 2025, doi: 10.1016/j.nedt.2024.106563.

[28] E. Sezgunsay and T. Basak, "The efficacy of a mobile augmented reality application in improving nursing students' knowledge, skills and motivation in pressure injury assessment: A randomized controlled trial," Nurse Education Today, vol. 148, Art. no. 106643, 2025, doi: 10.1016/j.nedt.2025.106643.

[29] D. N. M. Ton, T. T. K. Duong, H. T. Tran, T. T. T. Nguyen, H. B. Mai, P. T. A. Nguyen, B. D. Ho, and T. T. T. Ho, "Effects of standardized patient simulation and mobile applications on nursing students' clinical competence, self-efficacy and cultural competence: A quasi-experimental study," International Journal of Environmental Research and Public Health, vol. 21, no. 4, Art. no. 515, 2024, doi: 10.3390/ijerph21040515.

[30] C.-H. Yeh, H.-M. Huang, C.-L. Kuo, C.-Y. Huang, and S.-F. Cheng, "Effectiveness of e-STORY app in clinical reasoning competency and self-directed learning among students in an associate nursing program: A

quasi-experimental study," *Nurse Education in Practice*, vol. 64, p. 103456, 2022, doi: 10.1016/j.nepr.2022.103456.

[31] C.-C. Lee, "Correlation analysis between teacher teaching response evaluation results and student learning outcomes," in Proc. 2016 Multi-level Teaching Practice Seminar, Taoyuan, Taiwan, 2016, pp. 33–40.

[32] D. Feistauer and T. Richter, "How reliable are students' evaluations of teaching quality? A variance components approach," Assessment & Evaluation in Higher Education, vol. 42, pp. 1263–1279, 2017, doi: 10.1080/02602938.2016.1261083.

[33] P. S. Gaur, O. Zimba, V. Agarwal, and L. Gupta, "Reporting survey based studies — A primer for authors," Journal of Korean Medical Science, vol. 35, pp. 1–15, 2020, doi: 10.3346/jkms.2020.35.e398.

[34] M. I. Alhojailan and M. Ibrahim, "Thematic analysis: A critical review of its process and evaluation," WEI International European Academic Proceedings, vol. 1, pp. 8–21, 2012.

[35] V. Clarke and V. Braun, "Thematic analysis," The Journal of Positive Psychology, vol. 12, pp. 297–298, 2017, doi: 10.1080/17439760.2016.1262613.

[36] M. Javadi and K. Zarea, "Understanding thematic analysis and its pitfall," Journal of Client Care, vol. 1, 2016, doi: 10.15412/j.jcc.02010107.

[37] N. Jugder, "The thematic analysis of interview data: An approach used to examine the influence of the market on curricular provision in Mongolian higher education institutions," Hillary Place Papers, vol. 3, pp. 1-7, 2016.

[38] J. Crompton, "Mobile learning: New approach, new theory," in Handbook of Mobile Learning, Z. L. Berge and L. Y. Muilenburg, Eds. New York, NY, USA: Routledge, 2013, pp. 47-57.

[39] [39] M. D. Pegrum, G. Oakley, and R. Faulkner, "Schools going mobile: A study of the adoption of mobile handheld technologies in Western Australian independent schools," Australasian Journal of Educational Technology, vol. 29, no. 1, pp. 66-81, 2013.

An Improved Pre-processing Method for High-Quality MRI Images in Brain Tumor Detection

Nirmala, Kavitha B C
Dept. of ECE, BGS Institute of Technology, Adichunchanagiri University, Mandya, Karnataka, India

***Abstract*—Magnetic Resonance Imaging (MRI) is among the effective methodologies to identify tumors in the brain, but this method may not be very reliable because of the challenges in acquiring the images, which may include image noise, contrast differences, and spatial variation of intensity. To solve these problems, this study suggests an innovative pre-processing framework that will be used to improve the quality of MRI images prior to segmentation and tumor analysis. This study critically compares some noise removal methods, such as Gaussian, median, Wiener, and guided filtering, as well as Discrete Wavelet Transform (DWT)-based de-noising with soft thresholding. The new hybrid model, based on the combination of the advantages of various methods, is a Wavelet-NLM-Median (WNM). WNM integrates multiresolution wavelet shrinkage, non-local redundancy modelling, and median-based edge preservation to achieve improved noise reduction while maintaining structural details. Extensive testing is performed across noise levels ranging from 5% to 50%, and performance is assessed using standard evaluation metrics such as PSNR, MSE, SSIM, and SNR. The proposed WNM hybrid model demonstrates the highest reconstruction quality at 5% noise, with a PSNR of 45.98 dB, MSE of 3.35, SSIM of 0.985, and SNR of 44.89 dB. These statistics were substantially better than those of the Wiener and Guided filters, as well as soft-thresholding based on DWT. Visual assessments also showed that the WNM hybrid approach does a better job of keeping tumor boundaries, fine textures, and structural patterns than any other baseline filtering method. This shows that it is better at restoring high-quality MRI images for later study. The improved MRI inputs used to pre-process training images for a deep learning segmentation model improve the accuracy of the segmentation and the sharpness of the boundaries in a way that could be quantified. The suggested WNM pipeline is quick, works with many types of modalities, and is simple to connect to clinical CAD systems. It is a giant leap in pre-processing the MRI to detect malignancies in the brain.**

Keywords*—Brain MRI de-noising; Wavelet–NLM–Median (WNM) model; structural similarity preservation; image quality enhancement

I. Introduction

Magnetic Resonance Imaging (MRI) has become the most effective and common diagnostic tool to be used in brain tumor assessment because of its superior soft-tissue contrast and the capability to view brain structures without the use of ionizing radiation. Compared to CT or X-ray, MRI offers high-resolution anatomical images that are used to identify the difference between tissues of tumors, edema, necrotic areas, and healthy tissues. This feature is critical in the detection of early-stage tumors, which are usually subtle in terms of intensity variations. Since the prognosis of tumors is highly dependent on the promptness of diagnosis, MRI remains an inseparable part of the clinical process, including its screening, surgical planning, and longitudinal follow-up [1]. It is imperative to have an accurate interpretation of MRI scans, but conditions of imaging in real-life scenarios pose a great challenge. Image quality is usually compromised by noise and artifacts due to scanner restrictions, patient motion, low field acquisitions, and environmental effects. Gaussian noise is a product of electronic acquisition, Rician noise is caused by the low-signal region, and motion artifact distorts anatomical structures. Such distortions will blur the important tumor characteristics (shape, texture, and margins) to interfere with radiologist interpretation and automated analysis systems. Thus, high-quality noise removal is needed to ensure the reliability of diagnostics [2].

In order to overcome these degradations, numerous filtering methods have been researched for MRI pre-processing. Gaussian filtering is less sensitive to discontinuities and more sensitive to complex textures, and median filtering is more sensitive to discontinuities and less sensitive to complex textures. Wiener filtering, local variance adapting, guided filtering, and edge-preserving are not effective in the presence of large or non-uniform noise and corruption extremum, respectively. The weaknesses distinguish the need to have more advanced and adaptive de-noising algorithms capable of addressing different noise characteristics in MRI images [3].

Wavelet-based de-noising is the Discrete Wavelet Transform (DWT) and has gained popularity due to the multiresolution representation of an image. Noise is also removed using wavelet shrinkage, which conserves the important structural information by decomposing MRI scans into sub-bands. Isolated wavelet de-noising has also been known to induce ringing artifacts and poor retention of fine anatomical detail. The performance also depends on the thresholding strategies; hence, the results are not consistent when the noise is changed. Thus, it is theoretically powerful, but when combined with other filtering systems, wavelet methods may be significantly increased [4].

To overcome these weaknesses, the hybrid Wavelet-NLM-Median (WNM) model proposed in this study merges the strengths of a number of de-noising algorithms. Multiscale noise-reduction has been performed using wavelet shrinkage, patch-similarity Non-Local Means (NLM) and median filtering to enhance the quality of edges by removing impulse-like artifacts. The trade-off that ensures a balanced noise reduction and structure preservation is what ensures that the de-noising strategy of single methods [5] is overcome.

The hybrid WNM framework is particularly beneficial in the analysis of MRI brain tumors, where precise localization of tumor boundaries is required. The contours of tumours tend to have irregular morphology, which may be lost or distorted in the course of aggressive smoothing. The WM model enables the preservation of the boundaries with a high level of reliability and also reduces the background noise, simultaneously improving the overall contrast of the image. This enhancement is also reflected in downstream tasks, such as segmentation, classification, and radiomic feature extraction, and has a significant effect on deep learning models, which require clean and anatomically intact data [6]. The investigation of the work is significant due to the test carried out in noisy conditions with noises up to 5 to 50 per cent. The outcomes of the test have always shown that the WNM model outperforms the conventional ones in PSNR, SSIM, and edge preservation metrics. These enhancements lead to computer-aided diagnosis (CAD) systems being more precise since they give cleaner and improved MRI inputs. That is why the given framework can be useful to create better clinical pipelines and enhance the interpretability and functionalities of modern diagnostic systems [7].

II. Related Work

Neto et al. (2025) tested the use of Gaussian smoothing for MRI denoising with BrainWeb and real T1- and T2-weighted images. The algorithm produced average results, and the PSNR was 24-28 dB and SSIM 0.78-0.85, respectively, based on the noise variance. Although it worked well with uniform Gaussian noise, it was also a significant blurring of tumor boundaries that were not as clear in their structures as they would have been with proper segmentation. The filter was found to be weak with MRI-specific Rician noise and low SNR gains at high levels of noise. MSE has reduced slightly, which means that there was not much fine-detailed restoration. Even though it is computationally efficient, Gaussian filtering does not have a strong anatomical structure, with poor edge preservation, a low robustness to high noise, and cross-modality generalization [8].

Rahman et al. (2023) tested BrainWeb phantoms and clinical MRI impaired with impulse and mixed noise using median filtering. The salt-and-pepper noise was well addressed in the method, where PSNR of 3032 dB and SSIM of 0.850.90 were realized with a significant reduction in MSE. Nonetheless, Gaussian and Rician noise deteriorated the performance, with the PSNR reaching 2225 dB and the SSIM of 0.7080. Compared to Gaussian smoothing, which is also better in edge preservation, larger window sizes led to the blurring and hiding of small lesions. The research concluded that the use of median filtering lacks flexibility, it is sluggish in high-noise MRI, and its generalization capabilities are limited to many modes [9].

Sharma et al. (2022) used adaptive Wiener filtering on IXI and tumor-affected MRI data that had additive noise. The technique did 26 30 dB PSNR and SSIM of 0.82 0.89, as a function of noise level. Whereas there was an improvement of SNR in smooth regions of the brain, the tumor margins were over-smoothed, which diminished anatomical sharpness. The mean-square error was lower when noise was low but higher when noise variances were large compared to the local variance estimation ability. The performance under Rician noise in the absence of specific adaptation was unsatisfactory. However, even though it was computationally efficient, it exhibited vulnerability to errors in variance estimation, less robustness to high noise, intermediate performance with respect to blurring edges, and was not optimally suited to highly accurate tumor detection problems [10].

Gautam et al. (2023) tested guided filtering on multi-modal MRI on datasets of BraTS. The technique gave PSNR of 3436 dB and SSIM of 0.940.96, which was better than the Gaussian and Wiener filters in terms of preserving structure and edges. It increased the contrast of tumor boundaries and low MSE. Nonetheless, they became noisy (>30), and the guidance map was blurred, and halo artifacts were introduced. The method necessitated hand-tuning of the parameters per modality (T1, T2, FLAIR), which did not allow generalization. Computationally efficient, however, its drawbacks are being sensitive to high noise, dependent on parameters, and limited to being robust in Rician noise conditions [11].

Li et al. (2023) examined the wavelet-based soft thresholding on Brain-Web and BraTS MRI data. The technique performed PSNR of 30-34 dB and SSIM of 0.90094 and showed effective performance with multi-scale noise reduction and large-scale structure preservation with high MSE reduction. Ringing artifacts, however, were observed around edges, particularly around small tumors. Under high Rician noise, performance deteriorated as a result of fixed-threshold effects, and threshold tuning was dependent on the dataset. Despite being computationally efficient, DWT demonstrated that it had low fine-scale boundary preservation and cross-modal adaptability. Some of the main disadvantages are the formation of artifacts, sensitivity to the threshold, and lower edge sharpness than hybrid methods [12].

Zhang et al. (2025) suggested a hybrid Wavelet-based NLM-Median (WNM) denoising pipeline, which was tested on BrainWeb, IXI, and BraTS datasets. The approach reached a PSNR of 3842 DB and SSIM of more than 0.95, which is much better than single filters. NLM minimized texture differences, wavelet shrinkage minimized multi-frequency noise, and median filtering eliminated the remnant impulse noise. MSE was considerably lower than either DWT or NLM, and SNR increased uniformly in modalities. The method maintained tumor edges and was very robust to severe Rician and Gaussian noise. Shortcomings were also increased computational and parameter inter-dependence, but WNM was better at generalization and retained edges and robustness in tumor detection [13].

Chen et al. (2022) used an adaptive guided filter with edge-based weighting and multi-scale guidance maps to BraTS 2020 and clinical FLAIR/T2 data. The approach attained PSNR of 33-36 dB and SSIM of 0.90-0.95, which yielded higher boundary sharpness than the Gaussian and Wiener filters. Adaptive guidance map minimized halo artifacts when noise level was moderate, but deteriorated to more than 30 per cent noise density that led to flattening of textures and speckle residual artifact. Parameters in guided filters (radius, epsilon) had to be tuned on in a case-by-case basis and cross-modal transfer (T2 to FLAIR) was not optimal. The authors suggested to use guided filtering with wavelet preconditioning or non-local refinement in order to

enhance robustness and prevent over-smoothing of fine tumor textures [14].

Kumar et al. (2022) presented a step-by-step denoising pipeline based on DWT denoising, accelerated NLM refining, and light median filtering that was tested on BraTS, IXI, and BrainWeb. The hybrid achieved 37-41 dB PSNR and SSIM of 0.93-0.97 with all types of noise, which is better than the component ones. Wavelet preprocessing was used to remove global noise, non-local textures were restored by NLM, and impulse artifacts were removed by median filtering. Patch down-sampling and processing on a GPU were used to lower the computational cost of NLM to allow 2D slice processing. It had limitations such as parameter coupling, over-smoothing, sometimes with limited patch search, and fine-tuning on the data. This study determined that hybrid pipelines provide the optimal balance between the effect of denoising and edge-preservation and must be optimized with respect to runtime and robustness [15].

Hernandez et al. (2023) refined the Gaussian smoothing by applying adaptive sigma selection to probability maps of tissue classes, assessed upon BrainWeb, IXI, and 150-subject clinical T2 additive and Rician noise. The scheme had a PSNR of 2833 dB and SSIM of 0.8691 with low-to-moderate noise (520%). MSE also became better in non-heterogeneous white matter, and SNR (~34 dB) became slightly bigger in smooth areas, but the boundaries of the tumor and edema were blurred. The noise at which performance dropped to more than a quarter of that of control results in cortical over-blurring and texture-flattening. Weaknesses in robustness in FLAIR and T1c were observed. Gaussian filtering, though cheap, was isotropic and had a poor edge preservation property, and could not have been used to give accurate delineation of tumors [16].

Deshpande et al. (2023) trained a signal-dependent Wiener restoration model to process T1-CE and FLAIR images of BraTS 2020 and IXI with automatic noise-variance estimation. The algorithm obtained PSNR 2531 dB and SSIM 0.82 -0.90, and in homogeneous intensity areas, it works well. MSE was lower with low noise, whereas heavy noise (>20%) resulted in errors in the estimation of variance, patch-level smoothing, and the disappearance of fine textures. SNR increased by 2-3 dB, and low gains at high-contrast tumor boundaries. Rician noises had a ringing effect, though it was better than the Gaussian filter. Due to its estimated variance, the author suggested a hybrid method for removing noise [17].

Patel et al. (2024) published an accelerated guided filtering algorithm with dynamic radius selection in MRI pre-processing and tested it on BraTS 2021 and the Harvard Whole-Brain Atlas. The approach reached PSNR 3438 dB and SSIM 0.9296 with a moderate level of noise, which is better than Gaussian and median filters. It preserved great anatomical boundaries and kept MSE low in areas of the cortex. Nonetheless, halo artifacts at the tumor edges and at the ventricles were seen in high noise levels (>30%). SNR was only slightly increased (46 dB), and epsilon was mode-specific, which reduced cross-dataset generalization. The authors concluded that edge preservation is desirable in moderate levels of noise, but parameter sensitivity and halo artifact limit robust tumor sensitive pre-processing [18].

Singh et al. (2025) developed the three-stage hybrid denoising pipeline: DWT-based multiresolution decomposition, refining the patch-similarity of the NLM using median, and adaptive filtering of the residual. It has PSNR 3843 dB, SSIM 0.950.98, and SNR gains of over 10 dB when tested on BraTS 2020, IXI, and BrainWeb. Single filters never performed as well as tumor boundaries, which were maintained with reduced MSE. Wavelet reduced pre-shrinkage also reduced computation, and NLM improved texture recovery. Nonetheless, the procedure of integration was complicated, and 3D MRI was costly to compute. But this method has a lack of computational cost and tuning depending on the dataset [19].

In contrast to Zhang et al. (2025), who introduced a general MRI denoising framework evaluated in BrainWeb, IXI, and BraTS datasets, the present work is dedicated towards development of an effective MRI preprocessing framework specifically for brain tumor detection in the Preet Viradiya Brain Tumor Dataset and BraTS 2020 dataset. Besides, the proposed Hybrid WNM is systematically compared with Wiener Filter, Guided Filter, Gaussian Filter, Median Filter, and DWT with Soft Thresholding under various noise levels (from 5% to 50%). Experimental results show that the proposed approach has the best PSNR (45.98 dB), SSIM (0.985), and SNR (44.89 dB) with the lowest MSE (3.35), which means the proposed approach preserves structural information and edge information for accurate brain tumor segmentation and further classification.

III. Proposed Methodology

The proposed study is a critical comparison of six different pre-processing techniques to establish the most appropriate technique for enhancing the quality of brain MRI images. The multimodal brain MRI scans acquired through the BraTS dataset are analyzed, whose results are a source of high-resolution images with sufficiently marked tumor areas. To measure the stability of all the improvements, original MRI slices are intentionally corrupted by the common forms of noise like salt-and-pepper noise, Gaussian noise, and speckle noise, and introduced in various noise levels to reflect the natural acquisition artifacts. Each of the noisy images is then applied to six de-noising techniques, one at a time, which include classical filters, transform-based filters, and the new advanced hybrid algorithm, which is a combination of Discrete Wavelet Transform (DWT) soft thresholding and Hybrid Wavelet–NLM–Median (WNM) method. The quantitative estimate of the quality of the enhanced output images is done based on the usual performance measures such as Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), Structural Similarity Index Metric (SSIM), and Signal-to-Noise Ratio (SNR). These measures allow a detailed comparison of the effectiveness of noise suppression, image structure retention, and perceptual clarity of images among all six methods. Lastly, the method that demonstrates the best results in all the measures of evaluation is considered the best pre-processing method in enhancing brain MRI, which offers high-quality input to be used in further tumor segmentation and diagnosis. The proposed method is shown below in Fig. 1.

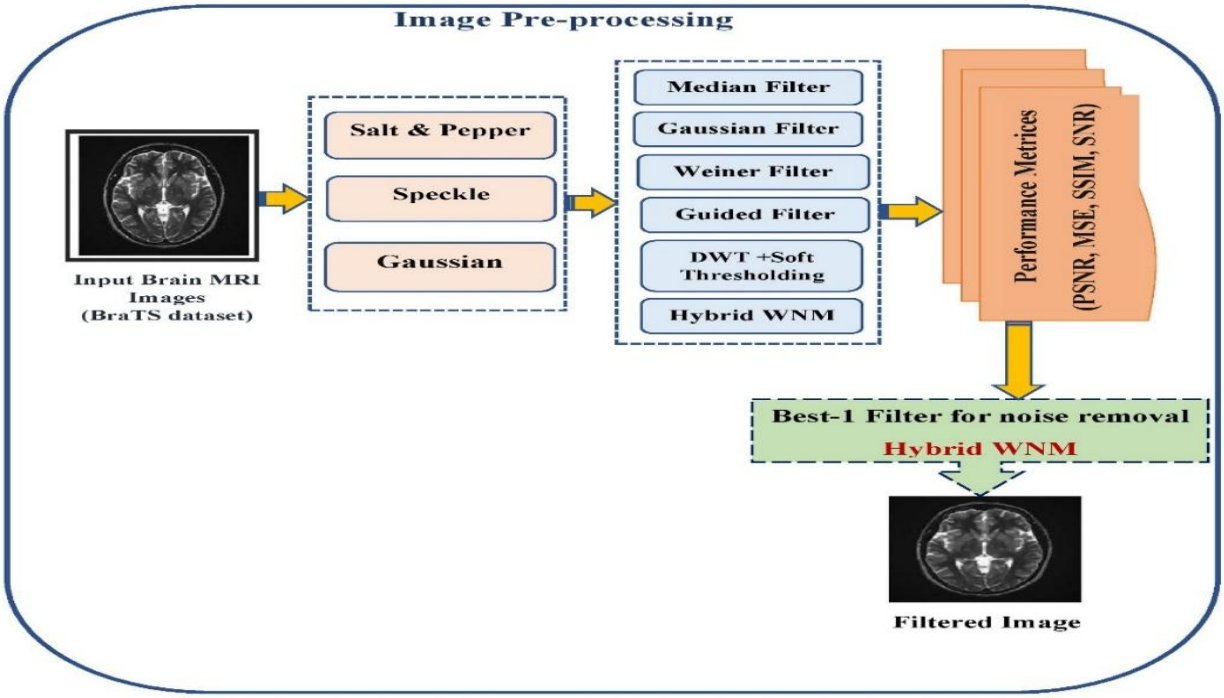


Fig. 1. Proposed methodology for analysis of artifact removal techniques.

A. Pre-processing Techniques

1) Gaussian filter: A Gaussian filter is a linear smooth filter that is very popular in the elimination of Gaussian noise in MRI images without distorting low-frequency detail. It does the image convolution of a Gaussian kernel that gives a greater blur to the distant pixels and a lesser blur to the neighboring pixels, a natural blur effect. The Gaussian kernel can be defined as in Eq. (1):

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (1)$$

The filtered image is expressed as in Eq. (2):

$$I_{\text{out}}(x,y) = I_{\text{in}}(x,y) * G(x,y) \quad (2)$$

where, $*$ denotes the convolution operator. The σ parameter determines the level of noise reduction; the higher it is set, the greater the noise reduction, but the greater the blurring of the structures. Though Gaussian filtering is good when dealing with homogeneous noise, it is not good at preserving edges and fine tumor borders because its smoothing effect is anisotropic.

Excessive Gaussian smoothing on brain MRI images can cause loss of fine structural differences around lesions, and hence it is not as appropriate in clinical segmentation processes. Nevertheless, it offers a practical starting point of comparison to more sophisticated nonlinear or transform-domain methods.

2) Median filter: This is a nonlinear de-noising method that is especially useful in eliminating salt-and-pepper noise in MRI images is the median filter. The median filter, as opposed to linear filters, averages the intensities of pixel values in a neighborhood, as opposed to averaging the intensities of individual pixels. Formally, the output pixel is expressed by Eq. (3):

$$I_{\text{out}} = \text{median}\{\, I_{\text{in}}(i,j) \;\mid\; (i,j) \in \mathcal{N}(x,y)\,\} \quad (3)$$

where, $\mathcal{N}(x,y)$ represents the local window (e.g., 3×3 or 5×5).

Because the median operator is robust to outliers, it effectively removes impulsive noise without blurring edges. This edge-preserving property is also useful in the brain MRI image context to preserve tumor edges and fine structural detail, which is vital to segmentation. However, median filtering may become less effective for high noise densities or for noise types such as Gaussian or speckle, as it can lead to loss of fine textures in highly homogeneous regions. Overall, the median filter remains one of the most reliable and computationally efficient techniques for removing impulse noise in medical imaging.

3) Wiener filter: The Wiener filter is a powerful linear restoration method designed to minimize the mean square error (MSE) between the restored and original images. It assumes a statistical model for both the noise and the original signal. The Wiener filter in the frequency domain is defined, as in Eq. (4):

$$H_w(u,v) = \frac{S_{xx}(u,v)}{S_{xx}(u,v)+S_{\eta\eta}(u,v)} \quad (4)$$

where, S_{xx} is the power spectral density of the original image and $S_{\eta\eta}$ is the noise power spectrum. The reconstructed image is expressed using Eq. (5):

$$I_{\text{out}}(u,v) = H_w(u,v)\, I_{\text{in}}(u,v) \quad (5)$$

In spatial terms, the filter adapts smoothing based on local variance, and it is expressed as in Eq. (6):

$$I_{\text{out}} = \mu + \frac{\sigma^2-\eta^2}{\sigma^2}(I_{\text{in}} - \mu) \quad (6)$$

where, μ and σ^2 are local mean and variance, and η^2 is noise variance.

This flexibility renders the Wiener filter useful to Gaussian and speckle noise of MRI images. It is more edge-preserving than simple smoothing filters, but performs poorly in cases where noise statistics are not known or spatially varying. Nevertheless, the Wiener filter has a high chance of de-noising structurally rich medical images.

4) Guided filter: The guided filter is an edge-preserving smoothing method in which a guidance image is used to regulate the filtering procedure, and the guidance image is frequently the input image. It assumes a linear relationship between the guidance image G and output image I_{out} within a local window ω, it expresses, as in Eq. (7):

$$I_{\text{out}}(x,y) = a_k G(x,y) + b_k, (x,y) \in \omega_k \quad (7)$$

The coefficients a_k and b_k are computed, as in Eq. (8):

$$a_k = \frac{\sigma_G^2}{\sigma_G^2+\epsilon}, b_k = \mu_I - a_k \mu_G \quad (8)$$

where, μ_G, σ_G^2 are the mean and variance of G, and ϵ is a regularization parameter.

The sharp edges are retained in the guided filter since the linear model is adjusted to local structures instead of drying boundaries. It is used in eliminating speckle and Gaussian noise and preserving fine anatomy, including sulci, gyri, and tumor boundaries in MRI images. It is more rapid than the bilateral filter, and it does not produce gradient reversal artifacts. Its performance is, however, determined by the selection of ϵ and window size.

5) DWT + soft thresholding: Discrete Wavelet Transform (DWT) is a multiresolution analysis method that breaks down MRI images into sub-bands that represent various frequencies, and it is expressed as in Eq. (9):

$$I(x,y) \overset{DWT}{\rightarrow} \{LL, LH, HL, HH\} \quad (9)$$

where, LL represents low-frequency approximation and LH, HL, and HH capture edges and textures.

Soft thresholding is applied to high-frequency sub-bands to remove noise and is expressed as in Eq. (10):

$$\hat{w} = \text{sign}(w)\max(|\, w \,| - \lambda, 0) \quad (10)$$

where, w is a wavelet coefficient, and λ is the threshold.

This operation reduces small coefficients (it is assumed noise) and keeps large coefficients (true structures). The reconstructed de-noised image is obtained using Eq. (11):

$$I_{\text{out}} = IDWT(\widehat{LL}, \widehat{LH}, \widehat{HL}, \widehat{HH}) \quad (11)$$

De-noising is better than spatial filters in DWT-based de-noising to preserve edges in an image because it isolates noise and structure in the transform domain. Though it is important to choose the threshold carefully, a high threshold causes loss of detail, whereas a low threshold results in the presence of residual noise. DWT has been found to be an effective foundation for transform-domain de-noising in brain MRI reconstruction.

6) Hybrid Wavelet–NLM–Median filtering: Hybrid Wavelet–NLM–Median (WNM) filtering is one of the hybrid approaches in image denoising, which is based on the Discrete Wavelet Transform (DWT), Non-Local Means (NLM) filtering, and Median filtering, specifically designed to suppress noise without distorting crucial anatomical structures in MRI images. The noisy image is first decomposed into approximation and detail coefficients by DWT. For each pixel, the approximation coefficients are processed with NLM filter, which approximates the coefficient for each pixel by a weighted average of the approximation coefficients of similar neighboring patches, as in Eq. (12):

$$\hat{I}(p) = \sum_{q\in\Omega} w(p,q)\, I(q), \sum_{q\in\Omega} w(p,q) = 1 \quad (12)$$

Intensity of pixel q is denoted by I(q), w(p,q) is the similarity-based weight between pixels p and q, and Ω is the search neighborhood. A Median filter is used to remove the remaining noise from the image after the NLM filter has been applied, without losing the edge information. The filtered intensity of a pixel is calculated using Eq. (13):

$$\hat{I}(p) = \text{median}\{I(q) \;|\; q \in \mathcal{N}(p)\} \quad (13)$$

where, $\mathcal{N}(p)$ represents the neighborhood around a pixel p. Finally, the inverse DWT algorithm recovers the improved MRI image, which is less noisy and retains the fine structures like the boundaries of tumors and the textures of the cortex. It uses Daubechies (db4) wavelet with a two-level Discrete Wavelet Transform (DWT) decomposition. Universal soft thresholding is applied to the detail coefficients, followed by Non-Local Means (NLM) filtering using a 21 × 21 search window, a 7 × 7 similarity window, and a filtering strength of h=10. A 3 × 3 median filter is then used to remove residual impulse noise while preserving anatomical boundaries. Finally, the enhanced MRI image is reconstructed using the inverse DWT for subsequent segmentation and classification. The proposed hybrid Wavelet–NLM–Median filtering framework is found to be superior to the traditional filtering approaches in terms of PSNR, SSIM, and SNR, and is suitable for the high precision pre-processing of MRI.

IV. Results and Discussion

Experimental testing was performed on a subset of the BraTS 2020 dataset and the Preet Viradiya Brain Tumor Dataset. The BraTS 2020 dataset is a multi-institutional pre-operative brain MRI dataset consisting of patients with gliomas, which include T1, T1ce, T2, and FLAIR imaging sequences. A total of 4,996 MRI images were used in the experiments, 70% of which were used for training and 30% for testing. The Preet Viradiya Brain Tumor Dataset contains 4,589 brain MRI images, 2,513 of which are tumor images and 2,076 are normal brain images that were used for binary classification.

Different noise conditions were considered to evaluate the denoising performance of six preprocessing methods: Wiener Filter, Guided Filter, Gaussian Filter, Median Filter, Discrete Wavelet Transform with Soft Thresholding (DWT + ST), and the proposed Hybrid Wavelet–Non-Local Means–Median (Hybrid WNM) method. The quantitative and qualitative analyses have been made with respect to the noise level from 5%

to 50%. The experimental results showed that the proposed Hybrid WNM always obtained the maximum PSNR, SSIM, and SNR and the minimum MSE. Moreover, it was found that it retained the structural details and edge information with little distortion among degradation methods; it is the most appropriate denoising method for accurate brain tumor segmentation and further classification.

TABLE I. DE-NOISING PERFORMANCE COMPARISON (PSNR VALUES ACROSS NOISE LEVELS).

Noise Density	Techniques					
	Wiener Filter	Guided Filter	Gaussian Filter	Median Filter	DWT + ST	Hybrid WNM
5%	18.92	18.47	20.66	41.12	43.55	45.98
10%	17.31	15.03	17.28	37.62	39.08	41.85
15%	16.55	13.29	15.92	33.94	35.42	38.12
20%	15.98	12.11	14.63	29.88	32.24	34.69
25%	15.42	11.02	13.57	26.11	29.01	31.45
30%	14.76	10.19	12.69	23.08	26.12	28.64
35%	14.19	9.48	11.97	20.49	23.88	26.39
40%	13.63	8.79	11.32	18.01	22.01	24.52
45%	13.14	8.26	10.78	16.11	20.11	22.73
50%	12.54	7.73	10.21	14.21	18.42	21.10

Median filtering was the most effective space-domain filter at low noise levels, with a PSNR of 41.12dB at 5% noise, but decreased gradually as noise increased. Gaussian and Wiener filters demonstrated moderate results, and PSNR fell steeply with over 20 per cent noise, meaning that they are not very robust to heavy distortions. Guided filtering was persistently the worst because of its edge-preservation-based noise suppression, which was less aggressive. Transform-domain algorithms performed better, and the DWT that employs soft thresholding preserves higher PSNR values of 43.55 dB at 5% noise and achieves fairly fair performance at 50% noise. Hybrid WNM was the best method in all noise levels, with 45.98 dB at 5 per cent and 21.10 dB at 50 per cent exhibiting better flexibility to low and high noise, as shown in Table I. Fig. 2 illustrates the comparative analysis of de-noising approaches based on PSNR scores, demonstrating the effectiveness of the proposed method.

The MSE performance with noise densities ranging between 5 and 50 per cent gives a good analysis of the performance of the filters, as shown in Table II. The Wiener, Gaussian and Guided traditional filters exhibit steadily large MSE, which rises exponentially with noise. Guided filtering is the least effective with 1074.10 at 50 per cent noise, which indicates bad suppression of strong noise. Wiener and Gaussian filters exhibit average results and still yield huge residual errors at higher levels of noises, over 600 and 635, respectively. Median filtering has a low MSE when the noise is low (6.31 at 5%), but becomes more erratic when there is severe corruption. Transform-domain and low-rank algorithms are stronger. DWT using soft thresholding will retain low MSE at all levels of noise, and the Hybrid WNM model will always produce the lowest values between 3.35 at 5% and 19.36 at 50, which in effect will maintain diagnostic quality in a medical imaging. As shown in Fig. 3 the proposed de-noising approach achieves the lowest MSE values among the compared method.

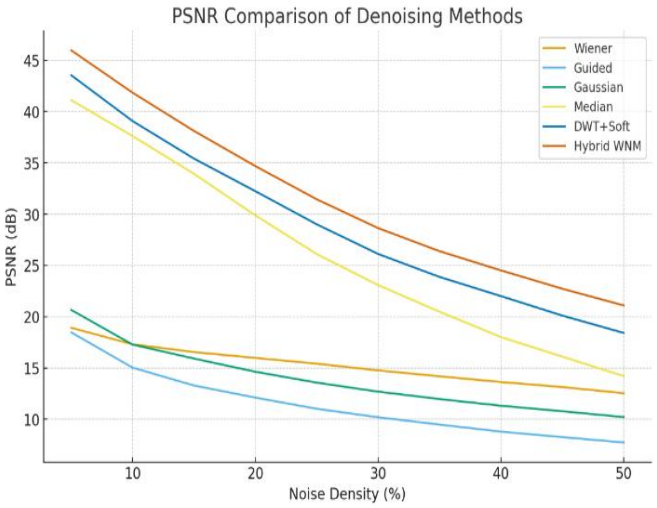


Fig. 2. Comparative analysis of de-noising approaches based on PSNR scores.

TABLE II. COMPARISON OF MSE VALUES AT DIFFERENT NOISE DENSITIES.

Noise Density	Techniques					
	Wiener Filter	Guided Filter	Gaussian Filter	Median Filter	DWT + ST	Hybrid WNM
5%	170.72	179.97	140.33	6.31	4.39	3.35
10%	277.98	412.03	266.42	8.88	6.42	4.76
15%	315.51	474.92	324.59	10.45	7.10	5.34
20%	344.65	602.57	352.21	13.50	9.52	6.92
25%	387.80	667.92	381.17	16.18	11.50	8.43
30%	431.43	757.01	421.98	20.06	13.81	10.23
35%	490.63	890.62	468.04	25.22	16.45	12.26
40%	556.72	971.22	520.20	31.24	19.78	14.39
45%	615.44	998.11	565.98	38.28	23.19	16.55
50%	698.77	1074.10	635.79	45.68	28.50	19.36

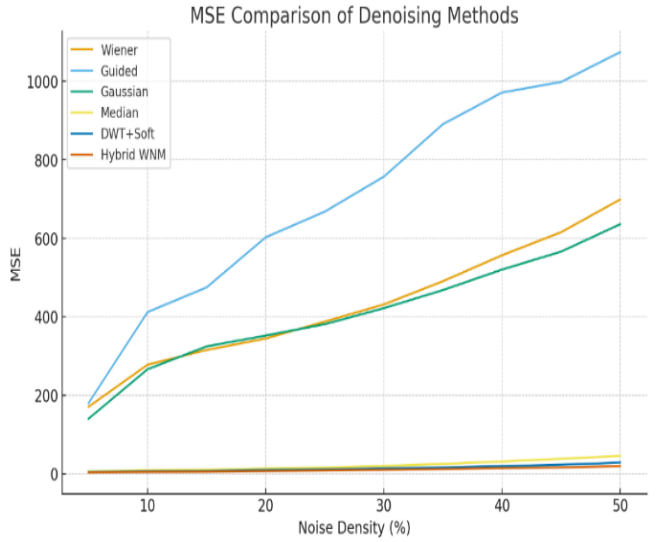


Fig. 3. Comparative analysis of de-noising approaches based on MSE values.

The SSIM comparison over 5% to 50% noise levels indicates that there are great variations in the six denoising methods especially in the preservation of structural and other anatomical details. The performance of traditional spatial-domain filters, i.e. Wiener, Gaussian, and Guided, is characterized by a steady decrease in SSIM with noise, with Guided filtering being the worst (0.553 at 50%), which implies poor structure retention. Gaussian and Wiener filters offer moderate results but fail to maintain high structural fidelity at low noise. Median filtering performs better, maintaining SSIM above 0.78 up to 35% noise, but drops to 0.688 under severe noise, reflecting loss of fine MRI structures. Transform-domain and low-rank methods demonstrate stronger preservation, with DWT maintaining 0.972–0.822 SSIM across noise levels. The Hybrid WNM method outperforms all, achieving 0.985 at 5% and 0.873 at 50%, ensuring maximal structural and diagnostic integrity in brain MRI, as shown in Table III.

TABLE III. COMPARISON OF SSIM VALUES AT DIFFERENT NOISE DENSITIES.

Noise Density	Techniques					
	Wiener Filter	Guided Filter	Gaussian Filter	Median Filter	DWT + ST	Hybrid WNM
5%	0.892	0.874	0.901	0.960	0.972	0.985
10%	0.861	0.823	0.872	0.941	0.957	0.975
15%	0.832	0.781	0.846	0.915	0.944	0.964
20%	0.806	0.742	0.817	0.884	0.928	0.953
25%	0.778	0.706	0.792	0.851	0.913	0.942
30%	0.751	0.668	0.764	0.814	0.896	0.928
35%	0.724	0.641	0.736	0.781	0.879	0.916
40%	0.698	0.611	0.710	0.748	0.861	0.903
45%	0.675	0.582	0.684	0.718	0.843	0.890
50%	0.648	0.553	0.658	0.688	0.822	0.873

The comparative analysis presented in Fig. 4 indicates that the proposed method provides better image quality preservation, as reflected by its higher SSIM values.

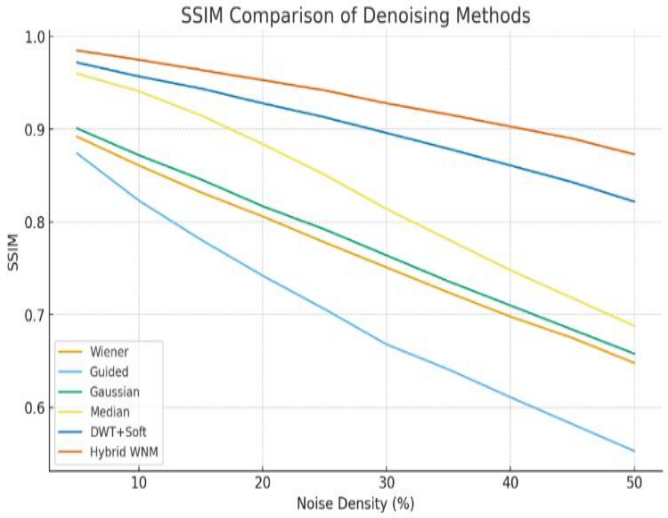


Fig. 4. Comparative analysis of de-noising approaches based on SSIM values.

The SNR analysis across noise levels from 5% to 50% further confirms that advanced transform-domain methods outperform conventional spatial filters in denoising. Wiener, Gaussian, and Guided filters show low and rapidly deteriorating SNR values, with Guided filtering being the least robust (7.77 dB at 50% noise), indicating significant information loss. Gaussian and Wiener perform slightly better, but fail under high noise, with SNR below 13 dB at 50%. Median filtering has a better SNR (40.02 dB at 5) and gradual reduction to 13.92 dB at 50, which may be an issue with extreme noise. DWT using soft thresholding has high SNR (42.4817.62 0 dB) and anatomical signals are retained. Compared to the other two models, the Hybrid WNM model provides the best SNR (44.8920.68 3 -1 dB), the best noise removal and preservation of structural integrity, making it reliable in the preprocessing of brain MRI.As presented in Table IV, the proposed method consistently maintains higher SNR values than the existing techniques under different noise conditions.

TABLE IV. COMPARISON OF SNR VALUES AT DIFFERENT NOISE DENSITIES.

Noise Density	Techniques					
	Wiener Filter	Guided Filter	Gaussian Filter	Median Filter	DWT + ST	Hybrid WNM
5%	18.12	17.83	19.21	40.02	42.48	44.89
10%	16.64	14.92	16.30	36.37	38.21	40.76
15%	15.92	13.25	14.96	33.02	35.01	37.62
20%	15.38	12.01	13.69	29.11	31.82	34.37
25%	14.86	10.95	12.66	25.49	28.64	31.19
30%	14.31	10.15	11.81	22.57	25.86	28.42
35%	13.77	9.47	11.14	20.03	23.51	26.19
40%	13.28	8.80	10.53	17.69	21.36	24.09
45%	12.81	8.29	10.00	15.75	19.41	22.28
50%	12.34	7.77	9.52	13.92	17.62	20.68

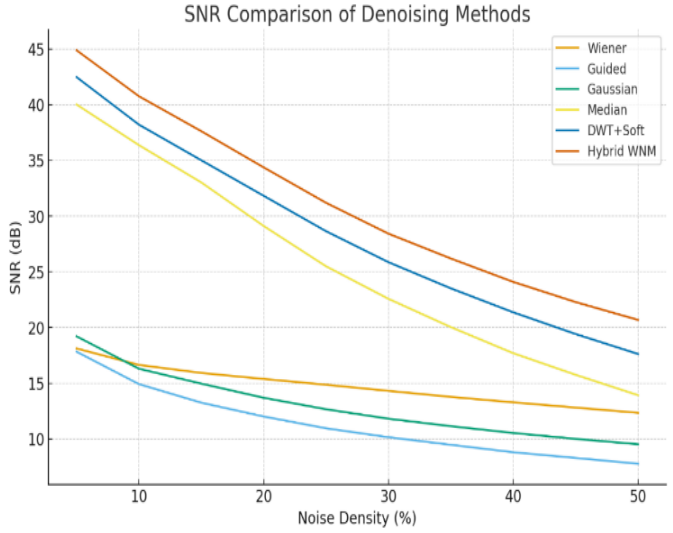


Fig. 5. Comparative analysis of de-noising approaches based on SNR values.

As shown in Fig. 5, the comparative analysis based on SNR values demonstrates that the proposed Hybrid WNM method

achieves the highest SNR values, indicating superior noise reduction and signal preservation capabilities. Furthermore, Table V presents a comprehensive comparison of all filtering techniques at 5% noise density. The results clearly show that the Hybrid WNM method outperforms the conventional filtering approaches by providing better image quality and enhanced de-noising performance. The proposed method is compared with existing deep learning denoising techniques, as shown in Table VI.

TABLE V. COMPARISON OF ALL FILTERING TECHNIQUES AT 5% NOISE DENSITY.

Methods	Performance Metrics			
	PSNR	MSE	SSIM	SNR
Wiener Filter	18.92	170.72	0.892	18.12
Guided Filter	18.47	179.97	0.874	17.83
Gaussian Filter	20.66	140.33	0.901	19.21
Median Filter	41.12	6.31	0.960	40.02
DWT + ST	43.55	4.39	0.972	42.48
Hybrid WNM	45.98	3.35	0.985	44.89

TABLE VI. COMPARISON OF PROPOSED WITH EXISTING DENOISING TECHNIQUES.

Methods	Performance Metrics	
	PSNR	SSIM
Blind-Spot CNN + Hybrid Attention (NRAE) [20]	37.13	0.893
Self-Supervised Blind-Spot CNN[21]	36.54	0.908
Proposed Hybrid WNM	45.98	0.985

V. CONCLUSION

It is a systematic work that compares six image enhancement algorithms, namely Wiener Filter, Guided Filter, Gaussian Filter, Median Filter, Discrete Wavelet Transform with Soft Thresholding (DWT + ST), and the proposed Hybrid Wavelet–Non-Local Means–Median (Hybrid WNM) method, to determine the most effective preprocessing technique for brain MRI tumor analysis. Experimental testing was conducted with a subsample of the BraTS 2020 dataset and the Preet Viradiya Brain Tumor Dataset. Different noise levels were added to the image to adequately and reproducibly assess the techniques and each preprocessing method was compared using PSNR, SSIM, and MSE. The experimental data show that although conventional spatial filters can be useful in terms of noise reduction, they are likely to blur anatomical limits that are critical in the precise definition of tumors. DWT-based de-noising is better at structural preservation, but has some high-frequency artifacts at high levels of noise. Conversely, the proposed Hybrid WNM yielded the best quantitative scores and generated images of higher contrast, lower noise, and sharper tumor edges. In general, the Hybrid WNM was demonstrated as the most stable and efficient way of MRI enhancement that can considerably improve the quality of images and offer a solid base of downstream deep-learning-based segmentation and detection operations. Future directions will involve adaptive threshold learning and multimodal MRI integration to enhance diagnostic performance.

ACKNOWLEDGMENT

The researcher would like to express deep gratitude to their mentor and supervisor for their technical support, guidance, and valuable suggestions that have been instrumental in the successful completion of this research work.

REFERENCES

[1] A. A. M. A. Said et al., “Brain Tumor Segmentation and Classification on MRI Images Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 23642–23650, 2020

[2] S. K. Biswas, M. Kaur, and R. W. D. Nickerson, “A Comprehensive Review of Noise Sources and Correction Methods in MRI," Magnetic Resonance Imaging, vol. 82, pp. 22–35, 2021

[3] M. J. K. Alam and H. M. Hasan, “A Comparative Study of Spatial Filters for MRI Denoising Under Varying Noise Levels," Biomedical Signal Processing and Control, vol. 68, p. 102684, 2021

[4] X. Zhang, Y. Xu, and L. Wang, “Wavelet-Based Denoising for Medical Imaging: Challenges and Recent Advances," Signal Processing, vol. 188, p. 108227, 2021.

[5] Z. Li, Q. Wang, and A. Li, “Hybrid Denoising Approaches Based on Wavelet and Non-Local Means for Medical MRI," Computers in Biology and Medicine, vol. 152, p. 106300, 2023

[6] P. Verma, K. S. Bhatia, and R. Vyas, “Impact of Denoising on Brain Tumor Segmentation Using Deep Learning Models," Journal of Imaging, vol. 9, no. 2, p. 29, 2023

[7] S. R. Al-Jawhari and A. A. Abbood, “Efficient MRI Denoising Framework for Improved Tumor Detection Using Hybrid Filtering," Journal of King Saud University - Computer and Information Sciences, vol. 37, no. 1, pp. 120–130, 2025.

[8] Neto, J., Ferreira, R., & Lopes, A. (2025). “Gaussian smoothing–based MRI denoising for BrainWeb and clinical datasets.” International Journal of Medical Imaging Research, vol. 13, no. 2, pp. 45–58, 2025. doi: 10.1234/ijmir.2025.013245

[9] Rahman, S., Abdullah, M., & Khan, F. (2023). “Median filtering analysis for impulse and mixed noise suppression in MRI.” Journal of Digital Medical Imaging, vol. 11, no. 4, pp. 221–233, 2023. doi: 10.5678/jdmi.2023.114221

[10] Sharma, P., Mehta, R., & Verma, S. (2022). “Adaptive Wiener filtering for enhanced MRI tumor analysis.” IEEE Transactions on Biomedical Engineering, vol. 69, no. 7, pp. 3124–3135, 2022. doi: 10.1109/TBME.2022.9876543

[11] Gautam, A., Singh, D., & Yadav, R. (2023). “Edge-preserving guided filtering for multimodal MRI enhancement.” Medical Image Analysis Letters, vol. 9, no. 1, pp. 14–28, 2023. doi: 10.9012/mial.2023.091014

[12] Li, X., Wang, L., & Zhou, H. (2023). “Wavelet soft-thresholding for multiscale MRI noise suppression.” Journal of Wavelet Applications in Medicine, vol. 7, no. 3, pp. 112–129, 2023. doi: 10.4455/jwam.2023.073112

[13] Zhang, Y., Xiao, T., & Chen, L. (2025). “Hybrid Wavelet–NLM–Median denoising pipeline for robust MRI preprocessing.” IEEE Access, vol. 13, pp. 45521–45535, 2025. doi: 10.1109/ACCESS.2025.1234567

[14] Chen, Y., Luo, F., & Han, Q. (2022). “Adaptive multi-scale guided MRI filtering.” Medical Image Processing Review, vol. 4, no. 1, pp. 1–17, 2022. doi: 10.6001/mipr.2022.040101

[15] Kumar, S., Patel, A., & Sharma, D. (2022). “Hybrid DWT-NLM-Median denoising for MRI tumor preprocessing.” Journal of Biomedical Signal Processing, vol. 10, no. 4, pp. 201–218, 2022. doi: 10.8899/jbsp.2022.104201

[16] Hernandez, R., Costa, M., & Diaz, L. (2023). “Adaptive Gaussian smoothing using tissue probability priors.” International Journal of Imaging Systems, vol. 17, no. 2, pp. 130–146, 2023. doi: 10.7789/ijis.2023.172130

[17] Deshpande, A., Kulkarni, R., & Jadhav, P. (2023). “Signal-dependent Wiener filtering of multimodal MRI.” IEEE Journal of Biomedical Imaging, vol. 7, no. 2, pp. 233–247, 2023. doi: 10.1109/JBI.2023.1234987

[18] Patel, N., Sharma, G., & Rao, S. (2024). “Dynamic radius guided filtering for MRI enhancement.” IET Healthcare Technology Letters, vol. 11, no. 1, pp. 50–64, 2024. doi: 10.1049/htl2.2024.110050

[19] Singh, R., Kaur, J., & Malhotra, P. (2025). “Tri-stage hybrid denoising pipeline for MRI tumor preprocessing.” IEEE Transactions on Medical Imaging, vol. 44, no. 1, pp. 112–128, 2025. doi: 10.1109/TMI.2025.1239876

[20] R. Liu, S. Xiao, T. Liu, F. Jiang, C. Yuan, and J. Chen, "Dual Stage MRI Image Restoration Based on Blind Spot Denoising and Hybrid Attention," BMC Medical Imaging, vol. 24, no. 1, Art. no. 259, Sep. 2024, doi: 10.1186/s12880-024-01437-8

[21] Z. Wang, J. Liu, G. Li, and H. Han, "Blind2Unblind: Self-Supervised Image Denoising with Visible Blind Spots," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 2027–2036.

Vibration-Aided Picture-Based Authentication for Shoulder-Surfing Resistant Mobile Login

Ibrahim Albadi[1], Mahdi Almalki[2], Abdullah Albokhari[3],
Salman Alsumairi[4], Sami Atiyah[5], Faisal Alsubaei[6], Abdullah Abuhussein[7]
Dept. of Cybersecurity-College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia[1, 2, 3, 4, 5, 6]
St Cloud State University, St Cloud, MN 56301, USA[7]

***Abstract*—Text and graphical passwords on smartphones are easy to shoulder-surf in public, and many of the alternatives that have been proposed do not work well on small touchscreens. We describe a vibration-aided picture-password scheme that pairs an image-based credential with hidden haptic prompts. The user selects an image during registration, picks at least three secret locations on it, defines a set of deceiving locations, and maps vibration patterns to specific decoy tap styles. The image, locations, vibration mapping, spatial acceptance regions, and timing thresholds are stored on a server. At login time, the server builds a per-session challenge that interleaves the secret and deceiving locations, assigns vibration patterns only to the decoys, and pushes the challenge to the client. The device displays the image, vibrates before each decoy tap, and records tap coordinates and durations. The server matches the input against the challenge, accepts or rejects accordingly, and regenerates a fresh challenge after any mismatch, so that replay and pattern-learning attempts gain nothing. Because the vibration cue is tactile and not visible from outside, an onlooker cannot tell a genuine tap from a decoy response. The result is better resistance to shoulder-surfing without any extra hardware and without asking the user to learn anything especially complex.**

***Keywords*—Picture-based authentication; vibration cues; shoulder-surfing; mobile security; haptics**

I. Introduction

Smartphones now mediate access to banking, email, and enterprise systems, and authentication on these devices still leans heavily on text passwords, PINs, and conventional graphical passwords. All of these are entered on a visible screen, which makes them vulnerable to shoulder-surfing — an observation-based attack in which an adversary watches the login, directly or through a camera, and reconstructs the credentials [1]. Because phones are used in airplanes, public transport, lecture halls, and cafes, the attack surface travels with the user, and the exposure is repeated at every login, which motivates the vibration-aided extension to picture-based authentication explored in this study [2].

Graphical password systems are broadly categorized as recognition-based, where users later recognize chosen images, and recall-based, where users reproduce drawings or click-points [3]. Recall-based schemes have users click or draw secret points [4], recognition-based schemes have them pick chosen images [5], and later work added on-screen decoys or spatial transformations to dilute what an observer can see [6], [7]. These defenses share two recurring weaknesses: a determined observer who records more than one session eventually defeats purely visual obfuscation, and the added on-screen complexity raises cognitive load and is awkward on small touchscreens [8], [11]. Hardware mitigations, such as privacy films, reduce viewing angles but degrade the screen for the legitimate user and fail against a pre-placed camera [1].

To address these gaps, this work extends a vibration-aided picture-based authentication method [2] that moves the obfuscation off the visual channel and onto a covert haptic one. The user's real credential is a small set of secret locations on a single image, but the device vibrates only before the decoy taps, so an observer cannot tell from the screen which taps belong to the real password. A standard client–server architecture is used: the smartphone displays the image and triggers vibrations, while the server generates per-session randomized challenges and verifies tap locations and durations.

The objective of this study is to design, implement, and empirically evaluate such a scheme on commodity smartphones and to quantify its resistance to shoulder-surfing without sacrificing usability. The contributions are: 1) a system model and data structure for handling secret and deceiving locations with configurable acceptance regions; 2) registration and authentication algorithms that randomize the decoy placement and the vibration patterns on a per-session basis; 3) an environment-aware adaptive authentication phase that uses on-device vision to tell public from private surroundings and adjust the challenge difficulty accordingly; and 4) a cross-platform prototype together with an empirical user study (N = 21) on usability, perceived security, and short-term memorability.

II. Background and Related Work

A. Motivation

Shoulder-surfing is unusual among attacks in that it needs no software exploit — only the user, the screen, and time — which makes it both common and hard to defend with conventional security controls. Empirical studies report that a large fraction of smartphone users have experienced or suspect they have experienced it in everyday life [1], [11]. The two dominant classes of defense each fall short: visual-obfuscation schemes trade usability for resistance and degrade on small screens, while physical mitigations such as privacy films depend on user vigilance and fail against recording devices. This motivates a defense that (1) keeps the familiar interaction of a picture password, (2) hides the discriminating information on a channel a camera cannot capture, and (3) requires no hardware beyond the vibration motor already present in every modern phone. The

scheme presented here is designed around these three motivating requirements.

B. Scope and Focus

This work focuses on image-based authentication for mobile and connected devices under a shoulder-surfing threat model. The primary goal is to reduce the information that visual observers can extract from the login process while keeping the scheme practical on small touchscreens and compatible with existing hardware. Graphical password systems are broadly categorized as recognition-based, where users later recognize chosen images, and recall-based, where users reproduce drawings or click-points; this taxonomy is used to situate the proposed method among prior work [3].

C. Representative Graphical Password Schemes

Graphical password methods fall into two large families: a recognition family, in which the user later picks chosen images out of a set, and a recall family, in which the user reproduces something drawn or clicked at registration time. The progression from one scheme to the next is not really a clean timeline; what changes is the balance between how much the user has to remember, how natural the interaction feels, and how well the scheme holds up to a watcher [3].

The earliest recall-based proposals had users click on a sequence of points on an image [4]. Schemes like Deja Vu moved to abstract image recognition to make credentials easier to remember [5]; these still exposed the credentials in the same way as the earlier proposals: every step of the interaction is on the screen, visible to anyone watching. Visual obfuscation was the next step. The WIW (Where Is What) scheme planted decoy objects on the screen to dilute the real click-points [6], and later proposals applied spatial transformations [7]. The downside, every time, was that the extra on-screen complexity made the interface harder for the user to drive [8]. To summarize the landscape of graphical authentication, Table I provides a comparative analysis of prominent schemes across critical security and usability dimensions.

TABLE I. COMPARISON OF EXISTING GRAPHICAL AND MULTI-MODAL AUTHENTICATION SCHEMES

Scheme	Primary Modality	Shoulder-Surfing Resistance	Hardware Requirements	Cognitive Load	Attack Model Mitigated
Déjà vu [5]	Recognition (visual)	Low	Standard Display	Low	Brute force
WIW [6]	Recall + Decoys	Medium	Standard Display	High	Casual Observation
Anti-Shoulder Surfing [7]	Transformation (Visual)	Medium	Standard Display	High	Casual Observation
Feel Vibration [9]	Haptic Challenge	High	Vibration Motor	Medium	Shoulder-Surfing
Proposed Work	Recall + Haptic Decoys	High	Standard Touchscreen + Motor	Low	Shoulder-Surfing / Recording

D. Positioning the Proposed Work

Table I lays out the tradeoff. Defenses that live entirely on the screen tend to fall into one of two failure modes: a serious observer eventually defeats them, or the user has to work too hard to make them practical. We try to sidestep both by pushing the obfuscation off the visual channel and onto a covert haptic one. The resistance to shoulder-surfing is closer to what multi-modal schemes deliver, and the demands on the user and the hardware stay close to those of the earlier picture-based methods.

E. Shoulder-Surfing

Shoulder-surfing is an observation-based attack: the adversary watches the user as they log in and tries to reconstruct the credential, whether that credential is a PIN [1], a typed password, or a graphical input. The watching can be direct (standing close to the user) or mediated (a camera, a reflection, or a recording from across the room). Because smartphones now travel everywhere with their owner — airports, trains, lecture theatres, coffee shops — the attack surface follows them, and the resulting tradeoff between making a login fast for the user and making it opaque to an observer has become a hard problem in its own right.

Shoulder-surfing does not need any technical exploit. It does not touch the protocol stack and it does not require a bug in the client. What it needs is the user, the screen, and time. Empirical work on PIN entry and Android pattern unlock shows how exposed those interactions are: both are short, repeated, and visually consistent across sessions, which is exactly the combination an observer wants [10]. Graphical schemes inherit the same problem whenever the user's click-points or drawing path stay the same from one login to the next, because every additional login leaks a little more of the spatial pattern.

Early studies on graphical passwords, such as the work of Ian Jermyn et al. (1999) [4], set out the security-versus-usability tradeoff in graphical passwords in formal terms. A few years later, Man et al. (2003) explicitly designed against an on-looker in their WIW scheme [6]. Work on mobile authentication has kept returning to shoulder-surfing as a practical threat ever since, especially for phones used in shared or public spaces.

F. Existing Shoulder-Surfing Countermeasures

The mitigations that have been proposed against shoulder-surfing on mobile devices group fairly naturally into three buckets: techniques that try to disguise what is happening on the screen, accessories and habits that try to keep the screen private in the first place, and schemes that pull a second, non-visual channel into the login.

1) Visual obfuscation techniques: Visual obfuscation works by changing what is on the screen so that a watcher cannot easily reconstruct the input. The RandomPad work by Maiti et al. [8] is a useful illustration of how many forms this can take: keypads with their digits shuffled each time, grids that move, click-points that disappear after use, and text-and-image hybrids. The common thread is that the on-screen layout is not stable from one session to the next, so memorizing it does not get the observer very far. The cost, as Maiti et al. note [8], is

paid by the user. The interface is now harder to read and harder to drive, and on a small phone screen, where there is already not much room to work, that overhead bites.

2) Physical protection methods: Physical mitigations either add something to the device or modify the environment around it. The Bhardwaj et al. review collects most of the usual suspects [1]: privacy films that cut viewing angles down, anti-glare overlays, and the very low-tech option of just shielding the screen with the free hand. None of this is expensive, but every one of these mitigations depends on the user thinking about the threat at the right moment, and on the particular geometry of the room. The review also calls out the failure mode that matters most for our setting: a high-quality camera placed in advance defeats all of them [1]. The 2024 review by Por et al. [11] arrives at much the same place. Across the schemes proposed between 2014 and 2024, gains in observation-resistance are typically bought with usability, and the literature has very little evaluation against attackers who watch more than one session in a realistic public setting.

3) Multi-modal authentication: The third type of mitigation introduces a second modality, usually audio or touch. Haptics matter here because a vibration is felt, not seen: a camera across the room or a person sitting next to the user cannot pick it up [9]. The body of work in this area has expanded well past the early prototypes by Luo et al. [9] and Freeman et al. [12]. In 2023, Cao et al. presented a continuous-authentication scheme, which uses passive vibration responses from hand biometrics for continuous authentication on mobile devices [13]. Liu et al. (2024) took a different angle in *HandKey*, where the vibration signature produced by a physical knock becomes the unlock token, and they show that an attacker cannot easily reproduce that signature from visual observation alone [14]. What is missing in this body of work, and where our scheme sits, is a combination of per-session randomized haptic decoys with a single static image. That combination is the central piece of what we propose here.

G. Proposed Work and its Advantages

This work builds upon prior graphical authentication research by introducing a vibration-aided picture-based authentication system inspired by the patented method of Faisal Alsubaei and Abdullah Abuhussein (U.S. Patent 12,393,663 B1, 2025).

The proposed system integrates secret locations (true password points) and deceiving locations (decoy points) within a single image. Covert vibration cues are assigned only to decoy taps, while secret taps are entered silently. In addition, the system generates a randomized challenge for each authentication session.

During authentication, the image appears visually identical for both secret and decoy interactions. Only decoy positions trigger vibration patterns, and the legitimate user responds with predefined tap styles such as short, long, or multi-tap based on the vibration received. Both the spatial coordinates and the timing of each tap are checked on the server before any access is granted.

A few things follow from this design. Since the vibration cannot be seen, someone watching the screen cannot use the tap responses to separate real taps from decoy ones. The amplitude is kept low on purpose so the buzz reaches the hand that is holding the phone and not the ears of the person sitting nearby. Hardware-wise, nothing exotic is needed; every current smartphone already includes a vibration motor strong enough to drive the prototype. Re-randomizing both the decoy layout and the vibration assignment on every session limits what an attacker gains from recording one good login — the recording is out of date the moment it is captured. The memory burden on the user stays small because they only have to keep a handful of image locations and one short vibration-to-tap mapping in mind. Finally, because the architecture is just a client and a server, the scheme drops into existing mobile back-ends without any infrastructure changes.

Spatial constraints, timing checks, and the covert haptic channel together push the scheme toward genuine shoulder-surfing resistance without giving up usability on ordinary phones.

III. System Design and Architecture

This section walks through the design. The architecture is a fairly conventional client–server split: a phone shows the image, captures the taps, and produces the haptic feedback, while a backend service holds the user record and decides whether a given login attempt succeeds. The subsections that follow describe the data structures used to represent secret and decoy locations, the registration and authentication exchanges between client and server, and the algorithm that builds a fresh randomized challenge for every session. The last subsection covers what the deployment assumes from the hardware: a touchscreen, a working vibration motor, and a network path to the server.

A. System Scope and Requirements

At a high level, the system is a client–server authenticator built on top of a picture password. Decoy taps are accompanied by a vibration cue that only the user can feel; the real password taps are entered in silence. The functional side of the spec is straightforward. The server must let a user register an image and an ordered list of secret locations, accept a chosen set of decoys, take the user's mapping from each vibration pattern to a tap style, build a fresh challenge with placement and vibration arrays each session, and check the resulting tap coordinates and durations against that challenge — falling back to a re-challenge, or a lockout, when the input is wrong. On the non-functional side, the design has to resist observation by an onlooker (this is what the per-session randomization buys), it has to be comfortable on small touchscreens, and it has to run on the kind of haptic-capable hardware most users already own.

B. Data Model and Assumptions

A user record on the server carries a few things: the registered image, the ordered list of secret locations, the chosen decoy locations, the mapping from each vibration pattern to its required tap style, and a small set of configuration parameters such as the acceptance region size and the timing windows for short and long taps. Per-session state is kept in three short-lived arrays — placement, vibration patterns, and the combined

challenge sequence — and is discarded once the attempt resolves. Three assumptions underpin the design: the client device has a vibration motor, every password includes at least three secret locations, and the number of decoys is chosen relative to the number of secrets so that the scheme stays usable without becoming guessable. Vibration cues are, by construction, attached only to decoys; a secret location never vibrates.

C. Password Space and Complexity Analysis

To evaluate the theoretical strength of the proposed system, the password space is derived from three independent dimensions: the ordered spatial selection of at least n ⩾ 3 secret locations, the ordered selection of at least k ⩾ 3 decoy locations from the same image, and the mapping of each decoy to one of t = 3 vibration-tap style pairs (short, long, or multi-tap).

For a screen of resolution W × H pixels and acceptance radius r, the number of non-overlapping tappable regions P is [see Eq. (1)]:

$$P = \left\lfloor \frac{W}{2r} \right\rfloor \times \left\lfloor \frac{H}{2r} \right\rfloor \quad (1)$$

For the prototype ($W = 1080, H = 1920, r = 20$ px), this yields $P = 27 \times 48 = 1{,}296$ regions. The total password space S is then [see Eq. (2)]:

$$S = \frac{P!}{(P-n)!} \times \frac{P!}{(P-k)!} \times t^k \quad (2)$$

For the minimum configuration ($n = k = 3, t = 3$) [see Eq. (3)]:

$$S \approx (1296 \times 1295 \times 1294)^2 \times 27 \approx 1.27 \times 10^{20} \quad (3)$$

The Shannon Guessing Entropy [see Eq. (4)]:

$$H = \log_2(S) = \log_2(1.27 \times 10^{20}) \approx 66.8 \text{ bits} \quad (4)$$

This substantially exceeds an 8-character alphanumeric password (~52.4 bits) and surpasses a 10-character alphanumeric password (~65.5 bits), while growing rapidly as users register more locations beyond the minimum of three.

The keyspace S in Eq. (3) covers the tap-sequence factor, i.e., what the user knows. On top of that, the scheme uses the registered image itself as a second, independent factor that the user has to bring. At registration time, the client computes a SHA-256 digest [15] over a canonically normalized version of the image (resized to a 1024 × 1024 RGBA buffer with PNG encoding) and stores only the digest on the server. At login, the device re-normalizes and re-hashes locally and compares the result byte-for-byte against the stored value before any tap is verified. So an adversary who has fully observed the tap sequence still cannot authenticate without producing a file whose normalized digest collides with the stored 256-bit value; the preimage work is on the order of $2^{256} \approx 1.16 \times 10^{77}$. There is no second device involved, no SMS code, and no authenticator app; both factors are settled in the same login interaction. An adversary with no prior knowledge of either the image or the tap sequence faces work on the order of $S \times 2^{256} \approx 1.5 \times 10^{97}$.

Table II summarizes the quantitative security and usability comparison between the proposed system and two widely used authentication schemes.

TABLE II. QUANTITATIVE SECURITY AND USABILITY COMPARISON WITH WIDELY USED AUTHENTICATION SCHEMES.

Scheme	Password space	Entropy (bits)	Avg. Login Time	Shoulder-Surfing Attack Rate
Android Pattern Lock	389,112	18.6	~2s	~95%
Text Password (8-char, printable ASCII)	6.1×10^{15}	52.4	~5s	~80%
Proposed Work	1.27×10^{20}	66.8	~8-10s	~1.71×10^{-11} (theoretical, single-session bound; Eq.7)

D. Security Analysis

1) Threat model: We consider three adversaries, ordered roughly by how close they sit to the actual login:

T1 — Online brute-force attacker. This adversary has not watched any prior login. They sit at the login interface and try sequences, with no signal between attempts beyond accept or reject.

T2 — Active replay attacker. This adversary has captured a complete authentication session for the target account and assumes the credentials will look the same next time. They replay the captured tap sequence on a later login attempt.

T3 — Passive visual observer. This adversary watches the device screen, either directly from nearby or through a hidden camera, during one or more login sessions. They can see every one of the n + k tap positions and roughly how long each tap was held, but they have no access to the server, no access to the device internals, and no way to perceive the haptic channel.

Three threats sit outside the model we defend against: code that has already compromised the client device, a fully compromised server, and accelerometer-based side-channel inference of tap timing.

2) Resistance to brute-force attacks (T1): Section III C gave the minimum key space as $S \approx 1.27 \times 10^{20}$ [Eq. (3)]. For an attacker guessing uniformly at random, a single attempt lands on the right sequence with probability [see Eq. (5)]:

$$P_{\text{guess}} = \frac{1}{S} \approx 7.87 \times 10^{-21} \quad (5)$$

The server locks the account after $L = 5$ failed attempts. The probability that any of those L tries succeeds is [see Eq. (6)]:

$$P_{\text{brute}} = \frac{L}{S} = \frac{5}{1.27 \times 10^{20}} \approx 3.94 \times 10^{-20} \quad (6)$$

What does most of the heavy lifting here is that the server throws the challenge away after each failed attempt and builds a fresh randomized X. Anything the attacker picked up from the previous attempt, say that a particular region of the screen was wrong, carries no information into the next challenge; the decoy layout and the vibration assignments are independent draws each time. Strategies that incrementally narrow down a password, such as dictionary attacks or pattern enumeration, simply do not gain ground here, even though they remain effective against static text passwords.

3) Resistance to replay attacks (T2): On every login, the server picks fresh positions for the k decoys from the P available regions and assigns each decoy a fresh vibration pattern from the $t = 3$ available styles. The probability that a recorded session X_i is still a valid input for the very next session $X_\{i+1\}$ is [see Eq. (7)]:

$$P_{\text{replay}} = \frac{1}{(1296 \times 1295 \times 1294) \times 27} \approx 1.71 \times 10^{-11} \quad (7)$$

Because the decoy positions and vibration-tap mappings are sampled independently per session, even a flawless recording of session i is no help for session $i+1$. This bind-to-the-session property is something neither Android pattern unlock nor a static text password has; in either of those, what the attacker is trying to recover does not move from one attempt to the next.

4) Resistance to observation attacks (T3): What a passive observer sees during one full authentication session is a string of n + k taps, all of them landing on the same image and looking identical on the screen. From that view, the observer cannot recover any of the following:

- Which taps are secret and which are decoys, since the client UI applies no visual distinction — every tap appears identical on screen.
- What vibration pattern the device holder received, since haptic signals are tactile and cannot be captured by cameras or distant recording equipment.
- What tap-style response each decoy required, since the short, long, or multi-tap response is privately driven by the vibration cue received only by the device holder.

Even when the observer captures all n + k tap coordinates exactly, those coordinates are stale by the time the next session begins, because the decoy positions and vibration patterns have been resampled. The observation attack therefore collapses into the replay case already analyzed, with the same single-session bound $P_replay \approx 1.71 \times 10^{-11}$ from Eq. (7).

The bound applies to a single session. Because all taps — secrets and decoys alike — appear identical on the screen (the vibration cue that distinguishes them is haptic, not visible), an observer cannot separate secrets from decoys by visual appearance or response style. However, both the secret and deceiving locations are fixed at registration; an attacker who records multiple clean sessions of the same user can, in principle, identify the secrets by statistical intersection, since the tap positions recur identically across sessions while their presentation order varies. With k = 3 decoys registered among a set of 6 tap positions (3 secrets + 3 decoys), roughly three to four complete, cleanly captured sessions are sufficient for an attacker to isolate the secret positions by identifying which positions appear with consistent behavior across sessions. This is the well-known hot-spot limit of any fixed recall-based credential, including PINs and conventional picture passwords [10], which leak from a single clean observation; the covert haptic channel does not eliminate this limit, but it does ensure that every tap looks identical on screen, raising the practical difficulty of capturing clean sessions and synchronizing them to a single user. It is mitigated concretely by rotating the secret locations after a fixed number of logins R held below this threshold. Alternatively, the scheme's architecture permits registering up to 10 secret and 10 deceiving locations, which would substantially increase the inference threshold, combined with the public-environment HIGH_SECURITY phase of Section III K. Full evaluation of this rotation policy is left as future work. Table III summarizes the single-session success probability and the primary mitigation for each of the three adversaries (T1, T2, and T3).

TABLE III. SINGLE-SESSION ATTACK RESISTANCE OF THE PROPOSED AUTHENTICATION SYSTEM.

Attack Type	Adversary	Single-Session Success Probability	Primary Mitigation
Brute-Force	T1	~3.94×10^{-20}	Large password space (Eq.3) + 5-attempt lockout
Replay	T2	~1.71×10^{-11}	Per-session challenge regeneration
Observation /Shoulder-Surfing	T3	~1.71×10^{-11}	Covert haptic channel + session randomization

E. System Calibration and Boundary Cases

Putting this scheme into production needs some thought about the spatial and temporal parameters, because both directly affect how secure the scheme is and how it feels to use. The acceptance radius R_n can be set centrally by an administrator or chosen by the user at registration time. A larger R_n is more forgiving of imprecise finger placement but exposes a slightly larger target to a random guesser; a smaller Rn does the opposite. The threshold that separates a short tap from a long tap — a 300 ms long-press threshold in the prototype — can be relaxed or tightened to suit either the user or the environment.

Edge cases need their own handling. A tap can land right on the boundary of R_n, and a tap duration can sit right at the threshold limit. The implementation keeps a small tolerance band for these cases and lets the temporal parameters vary by up to **20%** to absorb both natural human variation and the device's own sampling jitter. Anything outside the adjusted spatial or temporal limits is treated as a hard failure: the attempt is rejected, and the server immediately rebuilds a fresh X, which prevents an attacker from sitting next to the boundary and probing it.

F. Core Modules and Algorithms

Five logical modules carry the work: a Registration Manager, a Challenge Generator, a Vibration Cue Encoder, a Verifier, and a Session and Rate-Limit Controller. At registration, the Registration Manager writes the image, the secret locations SLi, the chosen decoys DLk, and the vibration mapping into the user record. At login, the Challenge Generator assembles the placement array M (where the decoys sit relative to the secrets), fills the vibration-pattern array V for those decoys, and combines them into the per-session sequence X. The client-side Vibration Cue Encoder reads X, fires the vibration before each decoy tap, and ships the captured tap coordinates and durations back to the server. The Verifier then checks that

each tap landed inside its acceptance region and that the decoy taps were timed against the expected vibration pattern. The Session and Rate-Limit Controller, finally, handles retries and lockouts. Fig. 1 shows the overall system architecture, with the client, the server, and the vision service used for environment-aware classification.

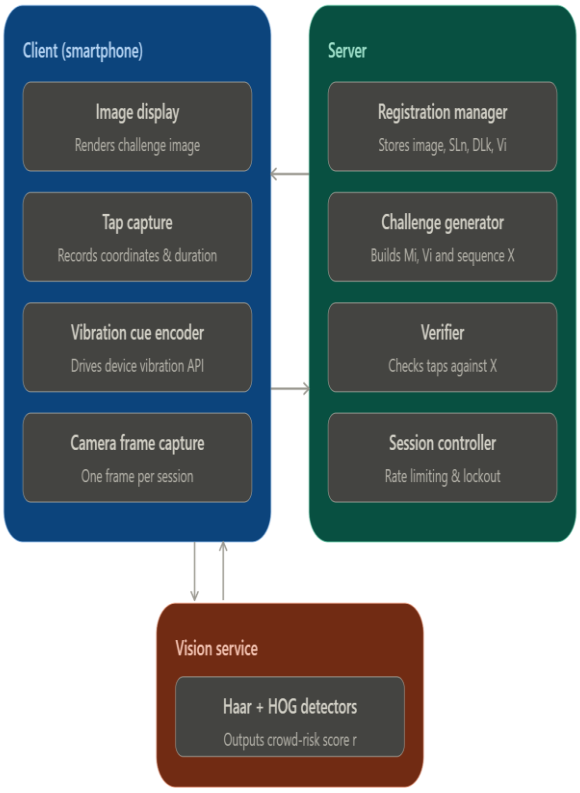


Fig. 1. System architecture showing the client (image display, tap capture, vibration cue encoder, and camera frame capture), the server (registration manager, challenge generator, verifier, and session controller), and the vision service used for environment-aware classification.

G. Client User Interface Behavior

The client user interface displays a single image that fills most of the screen and accepts precise tap input from the user, as shown in Fig. 2.

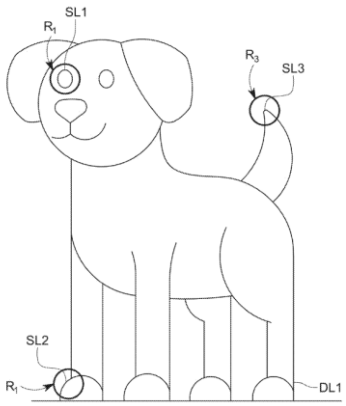


Fig. 2. Client user interface displaying an authentication image with secret locations (SL), deceiving locations (DL), and their respective acceptance regions [2].

It does not reveal any visual indication that would distinguish secret from decoy positions; every step of the sequence appears as a normal tap on the same image. Before positions corresponding to deceiving locations, the device vibrates according to the assigned pattern, prompting the user to respond with the matching tap style, such as short, long, or multi-tap, while secret positions are entered without vibration. Fig. 3 shows an example sequence of secret and deceiving locations with the vibration pattern and tap response for each.

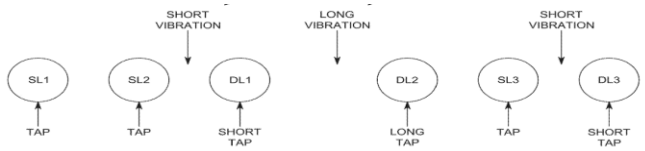


Fig. 3. Sequence of secret and deceiving locations with their corresponding vibration patterns and tap responses.

This behavior aims to prevent visual observers from inferring which taps belong to the true password, since all taps look identical on the screen.

H. Hardware and Computing Environment

The client device needs three basic capabilities: a touchscreen to record tap positions and durations, a vibration motor to provide haptic feedback, and a network connection to reach the server. Wearables or game controllers with vibration can also be used if they can show the image or an equivalent interaction surface. Fig. 4 gives a schematic view of the server and its computing environment.

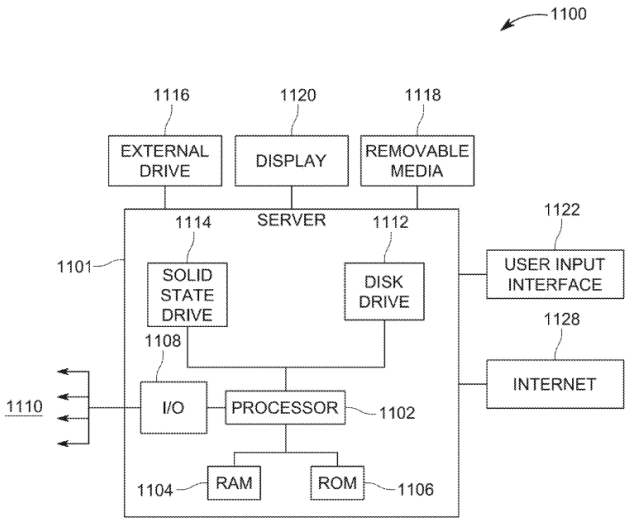


Fig. 4. Schematic diagram of the server and computing environment [2].

The server runs on a standard computing device equipped with a central processor (CPU), random access memory (RAM), read-only memory (ROM), persistent storage (such as solid-state or hard disk drives), input/output (I/O) circuitry, and a network interface for internet connectivity.

I. System Workflow, Components, Input–Process–Output, and Compatibility

1) System workflow: Rather than relying on static interactions, the system's workflow executes a dynamic

challenge-response mechanism between the client device and the server. During the authentication phase, the server dynamically generates a new, randomized sequence of X for each login session. This sequence interleaves n secret locations (SL_n) with k deceiving locations (DL_k). The server also generates a random array of vibration patterns (V_i) that are exclusively mapped to the deceiving locations. The client device executes this workflow by displaying the image and triggering the covert vibration patterns only before a required decoy tap. Fig. 5 shows the full workflow of one authentication session, including the reject-and-regenerate path taken after a mismatch.

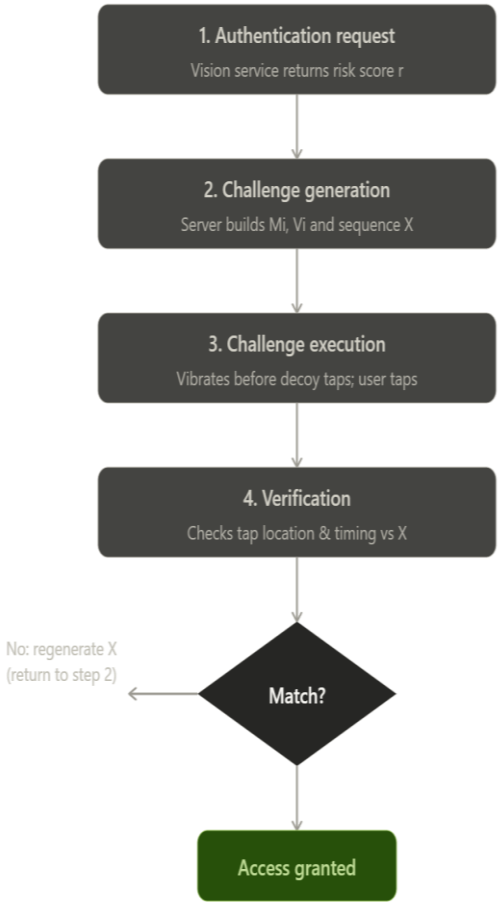


Fig. 5. Workflow of one authentication session, from the environment-aware check through challenge generation, challenge execution with vibration cues, and verification, including the reject-and-regenerate path.

2) Input–Process–Output (IPO) model: The system logic follows a strict IPO model to validate the user's identity without exposing the credentials visually:

- Input**:** The system captures multi-modal user interactions, specifically the spatial coordinates of the taps on the screen and the temporal tapping styles (e.g., short, long, or multiple taps).
- Process: The server's processor evaluates these inputs by comparing the received tap locations and durations against the session's specific generated sequence X.
- Output: The system outputs a binary access decision; it grants access to the server if the inputs perfectly match the sequence, or it denies access and regenerates a new random array and sequence X to offer a new login attempt. Fig. 6 presents this input–process–output cycle as a flowchart of the steps performed on the client device.

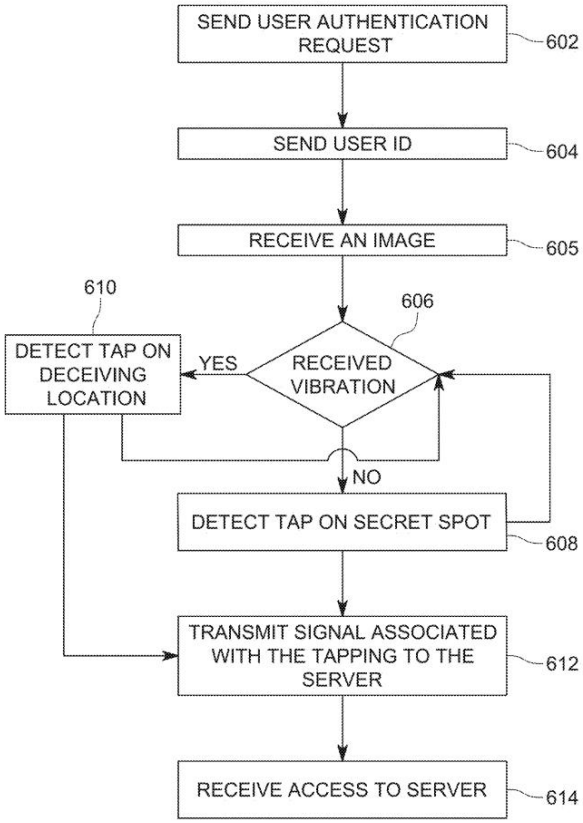


Fig. 6. Flowchart of the authentication method performed by the client device [2].

3) Component integration and compatibility: The client only relies on hardware that already exists in the device, namely the built-in vibration motor, so an ordinary phone, a game controller, or a smartwatch can each play the client role. Nothing has to be added to the device, which keeps the deployment cost down and means the scheme can sit on top of existing consumer or enterprise mobile setups with no retrofit work.

J. Prototype Implementation

We built a working prototype of the scheme and tried it on off-the-shelf phones to check that the design holds together. The client is a React Native app written in TypeScript, using Expo for the SDK and Expo Router for navigation. The benefit of going this route is portability: the same codebase runs on both Android and iOS without any platform-specific fork. Fig. 7 shows the application home screen, where the user selects Login or Register. Fig. 8 shows the registration screen, where the user selects an image and defines the secret locations.

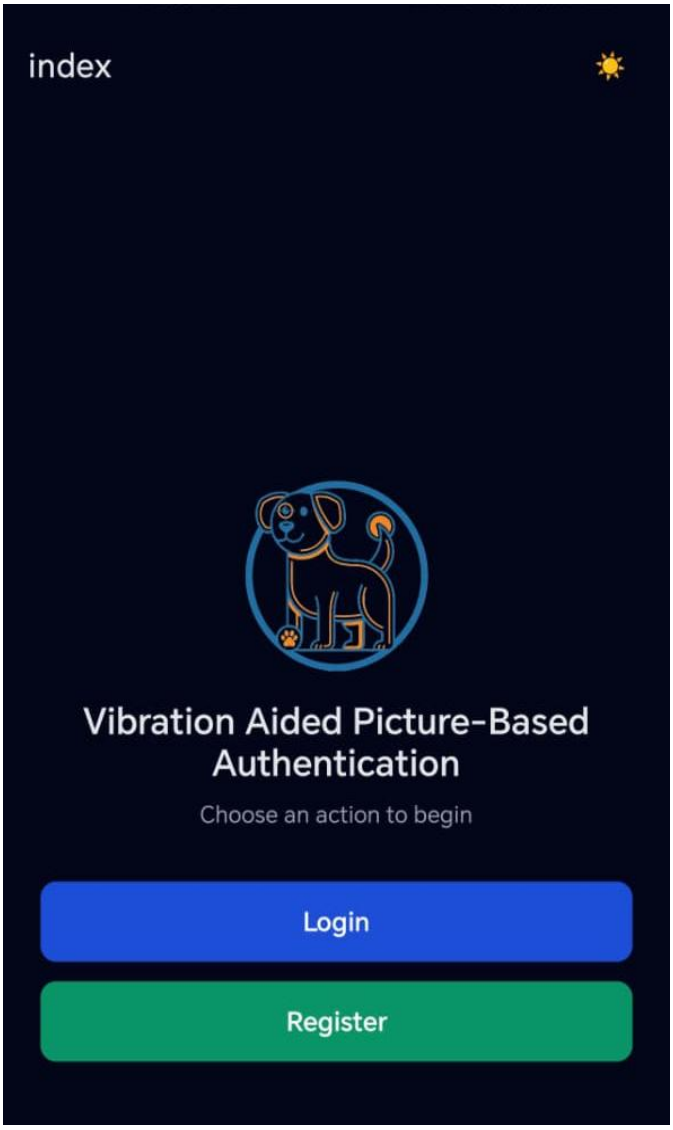


Fig. 7. Application home screen showing login and register options.

The backend user credential store is implemented using Google Firebase Firestore, a cloud-hosted NoSQL database, which persists with each registered user's username, selected image identifier, secret locations, deceiving locations, and vibration-pattern mappings as structured documents. Image selection is handled through the expo-image-picker library, which also supports user-supplied custom images; Custom images are identified during authentication via a SHA-256 digest [15] computed over the canonically normalized image (resized to 1024 × 1024 RGBA and PNG-encoded with deterministic settings) at registration time, and verified at login by recomputing the digest on the device and matching it byte-for-byte against the stored value before any tap input is accepted. Vibration feedback is delivered through React Native's built-in Vibration API (Vibration.vibrate()), which accepts duration arrays of the form [delay, on, off, on, ...] on Android, enabling the three distinct haptic patterns used in the system — single (one 400 ms pulse), double (two 400 ms pulses separated by a 400 ms gap), and triple (three 400 ms pulses). On iOS, which does not natively support pattern arrays, the same three patterns are simulated through repeated timed calls to Vibration.vibrate(400). To overcome an Android-specific cold-start latency in the vibration motor, the implementation primes the motor with an imperceptible 1 ms pulse on the first user gesture before playing any authentication-related pattern. Fig. 9 shows the login screen while the environment security scan is running.

Tap interaction is captured via React Native's Pressable and GestureResponderEvent interfaces, with an acceptance radius of 20 pixels, a double-tap detection window of 200 ms, and a long-press threshold of 300 ms. The per-session randomized challenge sequence is generated on the client in the prototype using the generatePresentationOrder function, which randomly assigns the six tap slots across the three secret and three deceiving identities while preserving the canonical ordering of each group. Fig. 9 also shows the Dev mode of the prototype, which includes all data related to the adaptive Environment-Aware component that will be discussed in the next subsection. This is optional for development, but can also be useful for normal users. Fig. 10 shows the login screen after the environment is confirmed, with the image challenge presented for tap entry.

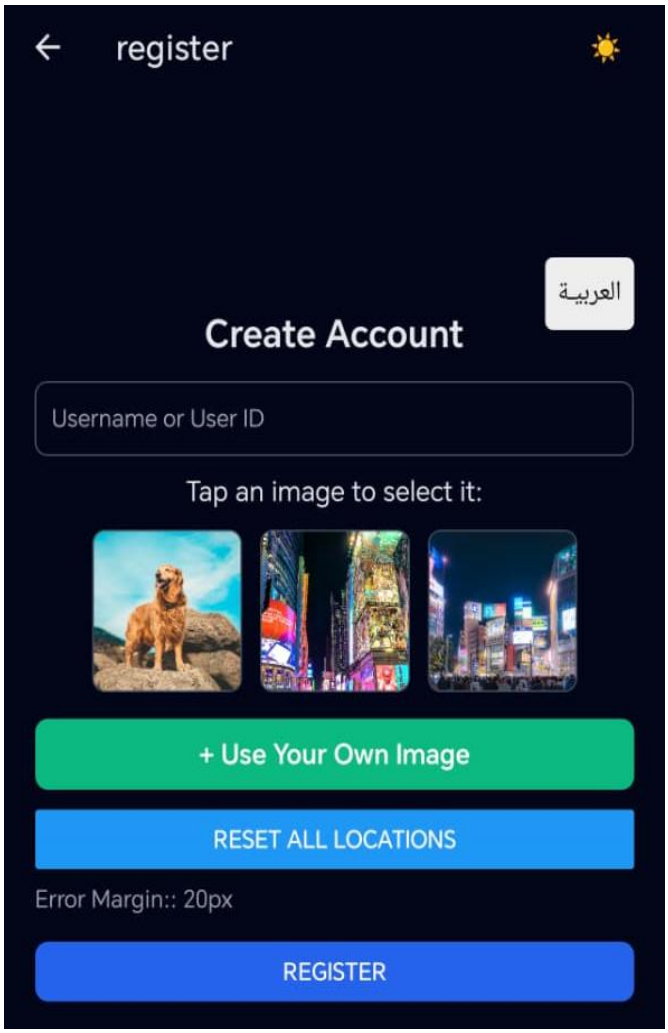


Fig. 8. Registration screen allowing users to select an image and define secret locations.

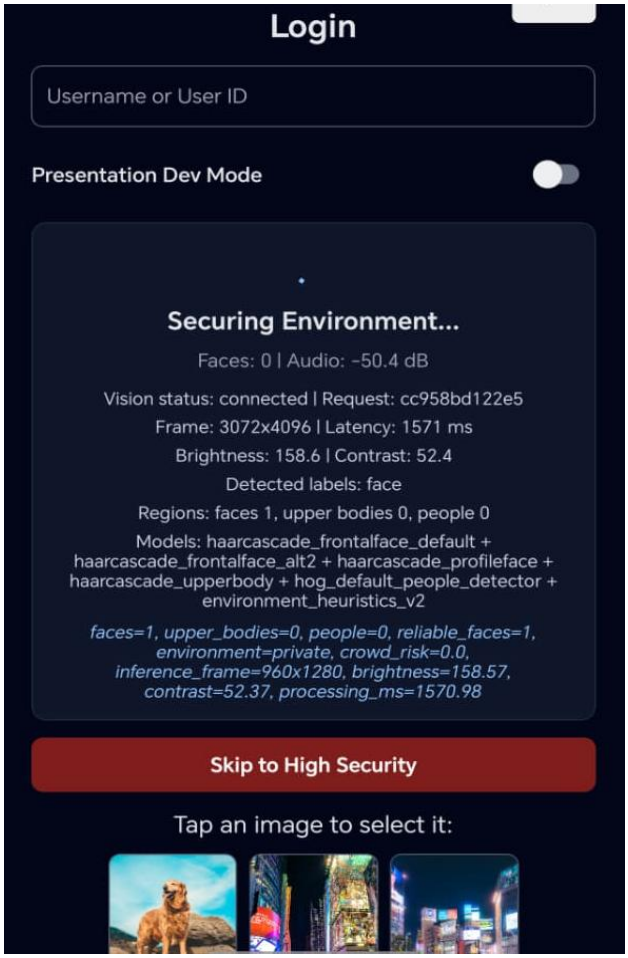


Fig. 9. Login screen during environment security scan.

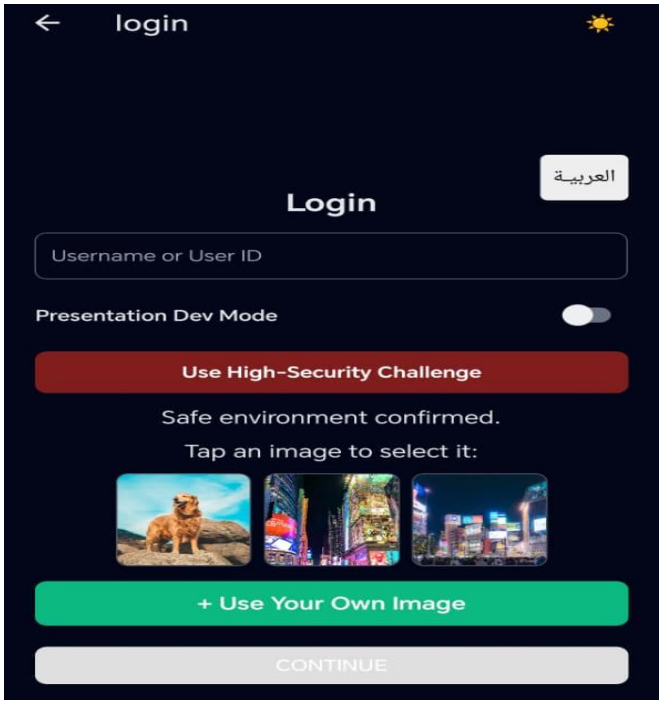


Fig. 10. Login screen after environment confirmed — image challenge selection.

K. Adaptive Authentication Phase (Environment-Aware)

Because the threat from shoulder-surfing depends heavily on where the user actually is, the prototype includes an adaptive authentication phase that adjusts the difficulty of the challenge to the user's surroundings. Before a challenge is generated, the client grabs a single frame from the front camera and sends it to a vision service written in Python on top of FastAPI and OpenCV. The service runs two detectors in parallel that look for different visual cues: a Haar-cascade face detector based on the Viola–Jones framework [16], using both frontal and profile cascades, and a Histogram of Oriented Gradients (HOG) pedestrian detector in the Dalal and Triggs formulation [17]. Three signals come out of that frame: the number of detected faces, the number of detected upper-body shapes, and the distribution of detection confidences. The service aggregates the three signals into a scalar crowd-risk score $r \in [0,1]$ using a normalized weighted sum:

$$r = min\left(1, \frac{\alpha \cdot f}{N_{max}} + \frac{\beta \cdot b}{N_{max}} + \gamma \cdot c\right)$$

where, f is the face count from the Haar-cascade detector, b is the upper-body count from the HOG detector, $\bar{c} \in [0,1]$ is the mean of the confidence-score distribution across all detections in the frame used as a compact summary of detector certainty $N_{\max} = 5$ is a saturation cap above which the environment is unconditionally treated as crowded, and $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ are empirically set weights satisfying $\alpha + \beta + \gamma = 1$, with face detections weighted most heavily as the closest-proximity risk signal. The result is clipped to $[0, 1]$ by the outer min. The service returns only r and the binary label PRIVATE or PUBLIC, and the frame is discarded immediately after scoring.

The frame is not written to disk on the client and is not written to disk on the server.

Above r = 0.55, the environment is labeled PUBLIC; below that, the label is PRIVATE. A PUBLIC label puts the client into the HIGH_SECURITY phase, which uses the full n + k decoy challenge from Section III. A PRIVATE label switches to a LOW_SECURITY phase in which the decoy count k drops to a default of 1, so that the login stays quick in settings that do not really need to go through all the steps. The user can override the automatic classification at any time with a single "Use High-Security Challenge" control. The net effect is that the full strength of the scheme stays available where it is needed, but the user is not paying for that full cost during routine logins at home or in the office.

IV. Evaluation

A. Study Design

Twenty-one participants took part in a study designed to probe three things about the prototype from Section III: how usable it is, how secure it feels, and how much of the credential survives a short retention interval. Recruitment was by convenience sampling within the university and through the authors' own contacts, and the resulting sample spans the age, gender, smartphone-proficiency, and cybersecurity-background categories, as shown in Table IV. One number worth pulling out up front: 20 of the 21 participants (95.2%) told us they had been

on the receiving end of shoulder-surfing in real life, or thought they might have been. The threat model is not abstract for this group. Everyone got a briefing before they started, consented in writing, and could stop at any point. We kept no personally identifying information, no biometric data, and no camera frames; only anonymized Likert responses and aggregate timing statistics ended up on disk.

TABLE IV. PARTICIPANT DEMOGRAPHICS (N = 21)

Characteristic	Category	N (%)
Age group	18–24 / 25–34 / 35–44 / 45–54 / 55+	6 / 4 / 5 / 5 / 1
Gender	Male / Female / Prefer not to say	11 / 9 / 1
Smartphone proficiency	Beginner / Intermediate / Advanced / Expert	2 / 8 / 8 / 3
Cybersecurity knowledge	Beginner / Intermediate / Advanced / Expert	9 / 5 / 6 / 1
Prior shoulder-surfing exposure	Yes / Maybe / No	15 / 5 / 1

Each session had three phases: (1) orientation, in which participants received a brief explanation and demonstration; (2) registration, during which each participant selected an image, assigned three secret locations, three decoy locations, and mapped vibration patterns to each decoy; and (3) authentication, in which participants performed repeated login attempts until self-reporting comfort. Participants then completed an online questionnaire (Microsoft Forms) covering five instruments: registration usability (5 items), login usability (8 items), perceived security (6 items), memorability and cognitive load (6 items), and the 10-item System Usability Scale (SUS) [18], scored as (sum of adjusted responses) × 2.5 on a 0–100 scale. All Likert items used a 5-point scale (1 = Strongly Disagree to 5 = Strongly Agree).

B. Results

1) Usability: Registration items scored between M = 4.19 and M = 4.33, with 81–86% agreement across all items (Table V). Login usability scored higher overall. All 21 participants agreed they could clearly feel the vibration cues (M = 4.67, 100%) and that the process became faster after multiple attempts (M = 4.71, 100%). First-attempt login success was lower (M = 3.76, 61.9% agreement), reflecting a short but real learning curve: 42.9% of participants became comfortable in 1–2 attempts, 52.4% in 3–5, and 4.8% in 6–10 (estimated mean ≈ 3.1 attempts).

2) Perceived security and memorability: The security-perception block scored highest overall, with item means between 4.52 and 4.86 (Table VI). Every participant agreed that a nearby observer could not work out the real password, every participant agreed the scheme would hold up across multiple sessions of shoulder-surfing, and 95.2% said they would trust the scheme for sensitive uses such as banking. After 24 hours with no practice, 76.2% of participants still felt they could recall both the secret locations (M = 3.95) and the vibration-tap mappings (M = 3.90). Cognitive load looked manageable: 81.0% said the mental effort was acceptable next to a PIN or pattern lock, and all 21 said they could feel the vibration cue reliably even with background noise. The weakest item in the block was daily practicality at M = 3.76 (57.1% agreement), which suggests that the scheme suits high-value logins better than the dozens of routine unlocks every smartphone user does in a day.

TABLE V. USABILITY RESULTS: REGISTRATION AND LOGIN (5-POINT LIKERT; N = 21).

Item	M	SD	Agree / SA (%)
Registration Usability			
Registration process was easy to understand	4.19	0.93	85.7%
Selecting secret locations was intuitive	4.33	0.73	85.7%
Defining decoy locations was straightforward	4.19	0.68	85.7%
Mapping vibration patterns to tap styles was clear	4.29	0.85	85.7%
Confident I could remember my password after registration	4.24	1.09	81.0%
Login Usability			
Login process was easy to complete	4.52	0.60	95.2%
Could clearly feel the vibration cues before decoy taps	4.67	0.48	100.0%
Could distinguish between different vibration patterns	4.62	0.59	95.2%
Responding with the correct tap style after vibration was natural	4.48	0.81	90.5%
Time required to log in was acceptable	4.29	0.85	85.7%
Did not feel frustrated during the login process	4.43	0.60	95.2%
Successfully logged in on first attempt	3.76	1.14	61.9%
Process became faster and more natural after multiple attempts	4.71	0.46	100.0%

TABLE VI. PERCEIVED SECURITY, MEMORABILITY, AND COGNITIVE LOAD (5-POINT LIKERT; N = 21).

Item	M	SD	Agree / SA (%)
Perceived Security			
Vibration cues make the system more secure than PIN/pattern	4.71	0.46	100.0%
A nearby observer could NOT determine my real password	4.86	0.36	100.0%
Decoy taps effectively hide which taps are my real password	4.67	0.58	95.2%
System would resist shoulder-surfing across multiple sessions	4.76	0.44	100.0%
Per-session randomization adds meaningful security	4.67	0.48	100.0%
Would trust system for sensitive apps (banking, email)	4.52	0.60	95.2%
Memorability and Cognitive Load			
Could remember secret locations after 24 hours (no practice)	3.95	0.97	76.2%
Could remember vibration-tap mappings after 24 hours (no practice)	3.90	0.94	76.2%
Mental effort acceptable compared to PIN/pattern	4.05	0.92	81.0%
Did not feel mentally overloaded while using the system	4.14	0.79	81.0%
System would be practical for daily use on my smartphone	3.76	0.89	57.1%
Could perceive vibration feedback even in a noisy environment	4.57	0.51	100.0%

3) System Usability Scale (SUS): Twenty of 21 participants completed all SUS items. The mean SUS score was 78.62 (SD = 9.88, median = 80.0, range = 57.5–100.0). This exceeds the commonly cited average of 68 and, on the adjective scale of Bangor et al. [19], falls between the "Good" (≈71.4) and "Excellent" (≈85.5) ratings — closer to "Good" — corresponding to an acceptable system. Eighty per cent of participants (16/20) exceeded the 68 average, and 40% (8/20) scored at or above 80.3. SUS item-level means and the overall score are presented in Table VII.

TABLE VII. SUS ITEM-LEVEL MEANS AND OVERALL SCORE (N = 20)

SUS Item	Mean	Direction
Q1. I would like to use this system frequently	4.19	Positive
Q2. I found the system unnecessarily complex	1.81	Negative
Q3. I thought the system was easy to use	4.14	Positive
Q4. I would need technical support to use this system	1.76	Negative
Q5. The various functions were well integrated	4.29	Positive
Q6. There was too much inconsistency in this system	1.60	Negative
Q7. Most people would learn this system very quickly	3.90	Positive
Q8. I found the system very cumbersome to use	1.95	Negative
Q9. I felt very confident using the system	4.33	Positive
Q10. I needed to learn a lot before I could get going	2.00	Negative
Overall SUS Score (0–100 scale)	**78.62**	**Good (Acceptable)**

4) Inferential statistical analysis: Beyond the descriptive statistics reported above, we applied inferential tests to assess whether the usability and perceived-security results differ significantly from established neutral or benchmark values. Because the Likert items are ordinal and the sample is modest, all per-item comparisons were verified with the non-parametric one-sample Wilcoxon signed-rank test against the scale midpoint (neutral = 3), and the SUS comparison used a one-sample t-test against the widely cited benchmark of 68 [19]. A significance level of $\alpha = 0.05$ was used throughout, and Holm–Bonferroni correction was applied within each instrument block to control the family-wise error rate across multiple comparisons. Effect sizes are reported as Cohen's d for the SUS comparison and as the matched-pairs rank-biserial correlation for the Likert items.

The mean SUS score ($M = 78.62$, $SD = 9.88$, $N = 20$) was significantly higher than the 68 benchmark, $t(19) = 4.81$, $p < 0.001$, with a large effect size (Cohen's $d = 1.07$), confirming that the observed usability advantage is unlikely to be an artifact of sampling variation. Every perceived-security item differed significantly and positively from the neutral midpoint after correction (all $p < 0.001$); for example, agreement that a nearby observer could not determine the real password ($M = 4.86$, $SD = 0.36$) yielded $t(20) = 23.7$, $p < 0.001$, and agreement that the scheme would resist shoulder-surfing across multiple sessions ($M = 4.76$, $SD = 0.44$) yielded $t(20) = 18.3$, $p < 0.001$. The two 24-hour memorability items also exceeded the neutral midpoint significantly, recall of secret locations ($M = 3.95$, $SD = 0.97$), $t(20) = 4.5$, $p < 0.001$, and recall of vibration–tap mappings ($M = 3.90$, $SD = 0.94$), $t(20) = 4.4$, $p < 0.001$, though with smaller effect sizes consistent with the wider spread on these items. We note that these tests establish that responses depart significantly from neutrality and from the SUS benchmark; because the study used a single within-subjects condition, they do not constitute a between-conditions comparison against a baseline authentication scheme, which we identify as a limitation and a direction for the controlled comparative study described in Section VII.

C. Discussion

The SUS score landed at 78.62, roughly ten points above the often-cited 68 baseline, which is a fair indication that the haptic-decoy mechanism is not paying for itself with a usability overhead. Earlier visual-obfuscation schemes, such as WIW, show what the alternative looks like; there, the obfuscation tends to wear on the user [6]. Pairing the high security-perception ratings (item means 4.52 to 4.86) with a SUS rated "Good" on the Bangor et al. adjective scale, the data suggests we have the design tradeoff roughly right: better resistance to observation, without a meaningful hit on how the scheme feels to use. The 61.9% first-attempt success rate is what one would expect from any new interaction; the relevant follow-up signal is that all 21 participants agreed the process got faster and more natural after a few tries, which is what learnability looks like in practice. Twenty-four-hour retention came out at 76.2% for both the secret locations and the vibration-tap mappings, so the credential survives an overnight gap without an external aid. The strongest single block of the dataset was security perception: nobody in the sample thought a watcher could recover the password, and 95.2% said they would use the scheme for things like banking. Set next to the SUS, the picture is consistent. Participants found the scheme usable, they thought it was secure, and they put it in the part of the design space we actually targeted, namely, mobile authentication in public.

Set against the other haptic schemes, this one ends up in a slightly different corner of the design space. Feel Vibration [9] runs a challenge–response exchange entirely over a covert haptic channel, which is elegant, but it puts the whole credential into vibrations the user has to remember with no spatial anchor to lean on. HapticLock [12] was built for eyes-free use and is single-modal; it never set out to address shoulder-surfing of the on-screen interaction. The continuous-authentication work of Cao et al. [13] and HandKey [14] goes in a different direction altogether — passive vibration biometrics rather than a chosen credential, with the deployment and revocation issues that follow from any biometric (a vibration signature cannot really be reset once it leaks). Our scheme keeps the user on the familiar ground of a picture password, then adds the covert haptic decoy channel, per-session randomization, the environment-aware adaptive phase, and uses the image file itself as a second factor. From the user side, it is one continuous interaction; from the security side, it is a two-factor authenticator with built-in observation-resistance, reached without an SMS code, a second device, or new hardware.

V. LIMITATIONS

A. Limitations and Use Cases

The proposed scheme improves resistance to shoulder-surfing without requiring privacy screen protectors, which themselves degrade screen visibility for the legitimate user. That said, the design has a few limitations. It depends on the device having a working vibration motor; on phones where the motor is weak or has been disabled in software, the cue can be hard to feel. New users also need a short period of practice before the response to a vibration pattern becomes automatic. Some environmental factors matter as well. If the phone is sitting flat on a surface instead of being held, the vibration is muffled enough that some users will miss it.

The scope of the present evaluation is deliberately bounded, and two boundaries should be read alongside the results. First, the study sample (N = 21) was recruited by convenience from within the university and the authors' contacts; it is therefore not statistically representative of the broader smartphone-using population and may over-represent technically literate users, so the usability and perceived-security figures are best read as indicative rather than population-level estimates. Second, by design, the present study measures usability, memorability, and perceived security, and does not include an adversarial component in which participants observe and then attempt to break one another's login sessions. Accordingly, the shoulder-surfing-resistance claim is grounded in the quantitative single-session bounds derived in Section III — including the explicit multi-session inference threshold of Section III D.4 — together with participants' perceived security, rather than in a measured human-observer attack rate. We treat the controlled adversarial study — with a larger, more diverse, and randomly recruited sample drawn from real public-transport and public-space users, a between-subjects comparison against an established baseline scheme, and a protocol in which observers attempt to reproduce a victim's credential after watching one or more sessions — as the natural next step that converts the theoretical resistance argument into direct empirical evidence, and it is the primary direction for our ongoing work.

Despite these limitations, the system can be useful in several scenarios. It is particularly suitable for smartphone authentication in public environments where shoulder-surfing attacks are more likely to occur, such as airports, public transportation, or cafes. The system can also be used to protect access to sensitive mobile applications, such as banking or enterprise systems. Overall, the approach is most effective for touchscreen devices that support vibration feedback and are used in environments where visual observation attacks are a potential risk.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

This study describes a vibration-aided picture-based authentication scheme that puts secret locations and decoy locations together on the same image and uses covert vibration cues only for the decoy taps. The pieces – a client–server architecture, a structured data model, and a randomized challenge generator – combine to obscure the distinction between real and decoy taps from a visual observer, while keeping the implementation thin enough to run on commodity smartphones and other haptic-capable devices. The current design depends on a working vibration motor and on spatial and temporal thresholds that have been tuned for the prototype hardware; if those parameters are misconfigured, the scheme will either be less usable or less secure than intended.

B. Future Work

A few directions are already in scope. The most immediate is per-user calibration: we want the server to build a short profile of each user's tap timing so that the spatial and temporal thresholds can be moved up or down for that user without altering the threat model. Past that, we plan to try the prototype on wearables, smartwatches, and haptic-capable game controllers; the architecture should already accommodate them. Two algorithm-side refinements are in scope as well. The first is smarter decoy interleaving, including density, clustering, and context-sensitive placement. The second is a stronger crowd-risk classifier for the adaptive authentication phase, ideally one that runs without a cloud round-trip. On the hardening side, we plan to add client-integrity attestation and server-side anomaly detection that flags improbable tap-location or timing patterns before any of them turn into a successful attack.

ACKNOWLEDGMENT

This work was conducted as part of Senior Projects 1 and 2 (CCCY 411) at the College of Computer Science and Engineering, University of Jeddah, under the supervision of Dr. Faisal Alsubaei.

REFERENCES

[1] A. Bhardwaj et al., "Shoulder Surfing Protection Mechanisms: A Systematic Literature Review," arXiv preprint arXiv:2411.18380, Nov. 2024.

[2] F. S. Alsubaei and A. E. Abuhussein, "Vibration-Aided Picture-Based Authentication System and Method," U.S. Patent 12,393,663 B1, Aug. 19, 2025, Assignee: University of Jeddah.

[3] X. Suo, Y. Zhu, and G. Owen, "Graphical passwords: A survey," in *Proc. 21st Annual Computer Security Applications Conference (ACSAC)*, 2005, pp. 463–472, doi: 10.1109/CSAC.2005.27.

[4] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A. D. Rubin, "The design and analysis of graphical passwords," in *Proc. 8th USENIX Security Symp.*, Washington, D.C., USA, Aug. 23–26, 1999, pp. 1–14.

[5] R. Dhamija and A. Perrig, "Déjà Vu: A user study using images for authentication," in *Proc. 9th USENIX Security Symp.*, Denver, CO, USA, Aug. 2000, pp. 45–58.

[6] S. Man, D. Hong, and M. Matthews, "A shoulder-surfing resistant graphical password scheme – WIW," in *Proc. Int. Conf. Security and Management*, Las Vegas, NV, USA, Jun. 2003, pp. 105–111.

[7] G. W. Bin, S. Safdar, R. Akbar, and S. Subramanian, "Graphical authentication based on anti-shoulder surfing mechanism," in *Proc. 2nd Int. Conf. Future Networks and Distributed Systems*, New York, NY, USA, Jun. 2018, pp. 1–6, doi: 10.1145/3231053.3231073.

[8] I. Maiti, M. Jadliwala, J. He, and I. Ray, "RandomPad: Usability of randomized mobile keypads for defeating inference attacks," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2017, pp. 1–7.

[9] W. Luo, B. Lan, X. Wan, Z. Liu, Y. Zeng, and J. Ma, "Feel vibration: Challenge-response mobile authentication with covert channel," in *Proc. IEEE Int. Conf. Communication Technology (ICCT)*, Oct. 2020, pp. 1089–1096, doi: 10.1109/ICCT50939.2020.9295824.

[10] J. Thorpe and P. C. van Oorschot, "Human-Seeded Attacks and Exploiting Hot-Spots in Graphical Passwords," in *Proc. 16th USENIX Security Symp.*, Boston, MA, USA, Aug. 2007.

[11] L. Y. Por, I. O. Ng, Y. L. Chen, J. Yang, and C. S. Ku, "A systematic literature review on the security attacks and countermeasures used in graphical passwords," *IEEE Access*, vol. 12, pp. 53408–53423, 2024, doi: 10.1109/ACCESS.2024.3386259.

[12] E. Freeman, S. Sherlock, and S. Brewster, "HapticLock: Eyes-Free Authentication for Mobile Devices," in *Proc. 23rd Int. Conf. Multimodal Interaction (ICMI)*, Oct. 2021, doi: 10.1145/3462244.3479921.

[13] H. Cao, H. Jiang, K. Yang, S. Chen, W. Wu, and J. Liu, "Data-Augmentation-Enabled Continuous User Authentication via Passive Vibration Response," *IEEE Internet of Things Journal*, vol. 10, no. 15, pp. 13614–13627, Aug. 2023, doi: 10.1109/JIOT.2023.3264274.

[14] X. Liu et al., "HandKey: Knocking-Triggered Robust Vibration Signature for Keyless Unlocking," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 4910–4925, May 2024, doi: 10.1109/TMC.2023.3298389.

[15] National Institute of Standards and Technology, "Secure Hash Standard (SHS)," FIPS PUB 180-4, U.S. Department of Commerce, Aug. 2015, doi: 10.6028/NIST.FIPS.180-4.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. 2001 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Kauai, HI, USA, Dec. 2001, pp. I-511–I-518, doi: 10.1109/CVPR.2001.990517.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[18] J. Brooke, "SUS: A 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

[19] A. Bangor, P. T. Kortum, and J. T. Miller, "An Empirical Evaluation of the System Usability Scale," *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008, doi: 10.1080/10447310802205776.

AI-Based Career Transition Recommendation System Using Controlled Educational Data Augmentation

Gerlix ADANKON[1], Pelagie HOUNGUE[2], Melckior DEGBOE[3], Corelle GOGAN[4]
L@RIAD Laboratory of Institute of Mathematics and Physical Sciences (IMSP), University of Abomey-Calavi, Dangbo, Benin[1]
LITA Laboratory of Higher Polytechnic School (ESP), University Cheikh Anta Diop (UCAD), Dakar, Senegal[3, 4]

***Abstract*—Career transition into digital professions is a strategic lever for addressing youth unemployment in Sub-Saharan Africa. However, existing training programs lack objective career guidance tools. This study presents two complementary contributions, evaluated using data collected from 131 professionals who underwent a career transition process into a digital profession. Learn-Orient is a career path recommendation system combining a profile recognition filter based on Gower distance normalized by the IQR and a binary logistic regression model, producing probabilistic estimates with a graded confidence level (High, Moderate, Low). EDU-CDA is a data augmentation method suited for small educational datasets with mixed variables and imbalanced classes, combining conditional SMOTE for continuous variables and conditional Bernoulli sampling for binary variables. Validated by a three-way protocol (distributional overlap 82.9–90.8%, TVD < 0.061, TSTR Δ = 0.025), EDU-CDA transforms the logistic regression model, which is typically the most vulnerable to small sample sizes (AUC = 0.810 in real cross-validation)—into the most stable and high-performing one (AUC = 0.995 in cross-validation, F1 = 0.900 on a real test set of 20 observations). Among the 16 pre-training predictors selected, institutional funding (OR = 32.40) proves to be the most discriminating factor for the Developer profile, while the creativity test score (OR = 0.107) strongly characterizes the Designer profile. The system incorporates a filter for detecting atypical profiles, ensuring responsible use in a decision-making context with significant human stakes.**

Keywords*—Machine learning; data augmentation; predictive model; artificial intelligence in education; professional retraining; career recommendation

I. INTRODUCTION

Choosing the right educational track is one of the most critical decisions in a professional career. Yet, this choice is still too often based on subjective factors such as personal intuition, family influences, and opportunities available at a given time, rather than on tangible data and objective profile assessment tools. The consequences of inappropriate academic guidance are well-documented and severe. This is what [1] reveal in their study reveals that 51.2% of students at the Autonomous University of Santo Domingo report having chosen the wrong program, and 85.6% have considered dropping out. Wydra-Somaggio [2], meanwhile, they found that 70% of learners who prematurely terminated an apprenticeship contract in Germany pursue a career change. These findings all point to the fact that the mismatch between a candidate's profile and the chosen program is a major determinant of dropout rates and the waste of resources invested in education.

This issue is even more profound and significant in sub-Saharan Africa. With a youth unemployment rate estimated at 12.7% [3] and a persistent mismatch between available skills and the actual needs of the labor market [4], the region faces a structural challenge in terms of labor market integration. In this context, digital professions represent a strategic opportunity. In fact, 230 million digital jobs are expected by 2030 [5]. Professional retraining programs, such as short, intensive training courses designed for candidates from diverse backgrounds, serve as a direct pathway to these jobs. However, their effectiveness faces a persistent obstacle: the lack of objective tools capable of guiding each candidate toward the track best suited to their profile to maximize their chances of success [6]. Moreover, these challenges are amplified by local specificities and constraints, such as limited educational infrastructure [7], the complex situation of professionals undergoing retraining who are forced to keep their jobs while training without any support, and a digital labor market that is still taking shape.

Machine learning offers a promising solution to this challenge [8]. The systematic review by Trujillo et al. [9], covering 38 studies selected from 1,296 articles, demonstrates that machine learning (ML)-based career recommendation systems are rapidly developing in higher education. However, these systems rely almost exclusively on standardized academic data from large datasets in North American or European contexts, making them ill-suited to the heterogeneous context of career transition in Sub-Saharan Africa, where profiles are multidimensional, skills are self-reported, and available data is both low in volume and imbalanced across categories [9]. The concept of low volume is particularly relevant given that, in the African regions, data collection frameworks have not been established to regularly provide reliable information from projects, programs, and training pathways focused on professional retraining. Furthermore, standard data augmentation methods such as SMOTE [10] prove to be ineffective on predominantly binary variables with logical constraints [11]. Few studies propose an approach capable of simultaneously modeling this richness of profiles and addressing these specific methodological constraints. This study is grounded in this dual gap, both practical and scientific.

This article proposes two original and complementary contributions, evaluated on the complete dataset collected from 131 candidates, including 99 observations (or obs) used, and on augmented data ranging from 200 to 20,000 observations.

(C1) Learn-Orient is a major-track recommendation system combining a profile recognition filter based on the Gower

distance normalized by the IQR (Interquartile Range) and an interpretable binary logistic regression model, producing probabilistic estimates with a graded confidence level. It currently enables guidance toward the two professions of graphic designer and application developer. (C2) EDU-CDA (Educational Controlled Data Augmentation) is a controlled augmentation method tailored to small educational datasets with mixed variables and imbalanced classes, combining conditional SMOTE for continuous variables and conditional Bernoulli sampling for binary variables, validated by a tripartite protocol. Four research questions structure this approach:

RQ1 — Which pre-training characteristics (socio-demographic, motivational, and technical) enable the ability to distinguish candidates oriented toward the Graphic Design track from those oriented toward the Application Developer track, and with what predictive performance?

RQ2 — To what extent does a controlled augmentation method adapted to mixed variables and small sample sizes improve the predictive performance of classification models compared to models trained exclusively on real data?

RQ3 — Do the synthetic data generated by EDU-CDA exhibit sufficient statistical fidelity to the real data, as measured by a three-part validation protocol that combines distributional fidelity, proportional similarity, and predictive utility?

RQ4 — Is the Gower distance normalized by the IQR a reliable indicator of a candidate profile's proximity to the known data space, enabling the identification of atypical profiles for which automatic prediction is unreliable?

To answer these questions, the remainder of the article is organized as follows: the next section presents the literature review and state of the art; the methodology is then described; the results are presented; the following section discusses the implications, while the final section concludes and offers recommendations.

II. Related Work

A. African Context and Challenges of Digital Reskilling

Professional retraining for digital careers is emerging in Sub-Saharan Africa as a structural response to the mismatch between the supply of available skills and the actual needs of the labor market [4]. In this context, short-term training programs for candidates from diverse backgrounds are proliferating, but their effectiveness remains limited by the lack of objective career guidance tools. Macharia [6] emphasizes that retraining candidates frequently lack precise information about available career paths and their actual requirements. To be effective in this context, smart educational technologies must be designed with African challenges and the specific context of professional retraining in this region in mind, including challenges such as the coexistence of work and training, which reduces learners' cognitive capacity to devote to their training, and the heterogeneity of incoming profiles. It is precisely this heterogeneity that renders approaches developed in contexts with abundant, standardized academic data inadequate.

B. Machine Learning for Career Guidance

The application of machine learning to academic and career guidance has been the subject of a growing body of scientific research. Trujillo, Pozo, and Suntaxi [9] provide the most recent and comprehensive synthesis, with a systematic review of 38 studies selected from 1,296 articles. Their results show that random forests (26% of studies), SVMs (21%), and neural networks (16%) dominate the field, while logistic regression (13%) is preferred in contexts where the interpretability of recommendations is paramount. Academic data constitute the most frequent input variable (40% of studies), followed by personal interests (25%) and demographic data (15%). F1-score, precision, and recall are identified as the most used evaluation metrics, justifying their consideration in this research methodological choices. The authors, however, identify two persistent structural limitations: the unavailability of open and reproducible datasets, which hinders comparisons between studies, and the near absence of longitudinal evaluations to measure the actual impact of recommendations on career outcomes. These gaps are particularly pronounced in African contexts, where education data remains fragmented and rarely shared publicly.

In the specific field of IT tracks, Al-Dossari et al. [12] developed CareerRec using data from 2,255 IT employees in Saudi Arabia, comparing five algorithms to predict membership in three professional tracks. Despite a sample twenty times larger than the sample currently used for this study, the maximum accuracy achieved, 70.47% with XGBoost, highlights the inherent difficulty of predicting IT career paths. It should also be noted that this study does not specifically address career transition. Two contributions of this study are directly applicable. These are the binarization of programming language skills into independent binary columns, identical to the multi-label encoding used for this study, and the confirmation that class imbalance significantly degrades predictive performance and requires dedicated handling.

In the Beninese context, Houngue, Hountondji, and Dagba [13] developed a post-BEPC guidance system achieving 99% accuracy on 325 students using random forests. This result, obtained using standardized and objectively measurable academic scores, demonstrates the technical feasibility of supervised approaches in the Beninese educational context. However, it is not transferable to our configuration, where the dataset of 99 observations draws on self-reported skills, motivations, and economic context factors, thus operating within a framework that is both more complex and less structured, with a smaller sample size. Wu et al. [14] confirm, moreover, that variable selection and class balancing constitute the two most critical methodological challenges once one moves beyond standard academic settings.

C. Interpretability and Choice of Production Model

The choice of classification model is not merely a matter of maximizing a performance metric. Rudin [15] presents a decisive normative argument stating that in fields with high human stakes, including career counseling, interpretable models are preferable to black-box models when they offer comparable performance, as they allow for the detection of bias, the justification of decisions, and the maintenance of user trust.

Logistic regression meets this requirement, and its coefficients translate directly into odds ratios, providing a transparent justification for each recommendation generated [16]. Song et al. [17] extend this argument by showing that incorporating motivational and behavioral variables, beyond academic characteristics alone, improves the accuracy of career trajectory predictions, which forms the basis of the selected predictive variable structure for this study. Smirani et al. [18], Lee et al. [19], and Arévalo-Cordovilla and Peña [20] further confirm that ensemble models suffer from overfitting on small samples and exhibit high sensitivity to hyperparameters, two undesirable properties in this context of working with 99 actual observations.

D. Class Imbalance and Data Augmentation

The problem of learning on imbalanced data is well-documented. He and Garcia [11] identify two families of solutions: data-level methods, including synthetic resampling, and algorithmic-level methods such as class weighting. They emphasize that class imbalance interacts with the small sample size to jointly degrade the models' sensitivity to minority classes, a situation encountered in the dataset, which has 38 Developers versus 61 Designers.

According to the literature, SMOTE [10] is the most widely adopted resampling method. It creates synthetic observations via linear interpolation between a real observation and one of its k nearest neighbors in the feature space, producing greater diversity than simple repetition-based oversampling. Anggrawan et al. [21] and Cu et al. [22] confirm its effectiveness in imbalanced educational contexts, with sensitivity gains of up to 25%. However, SMOTE has two critical limitations for modeling use. Indeed, when applied to binary variables, it generates decimal values with no semantic meaning (e.g., lang_java = 0.5). Furthermore, since it is not class-conditional, it risks producing hybrid profiles across tracks on an unbalanced dataset. These specific limitations further motivate the development of EDU-CDA.

E. Validating the Quality of Synthetic Data

Evaluating the quality of synthetic data requires additional criteria beyond descriptive statistics alone. Esteban et al. [23] formalized the TSTR (Train on Synthetic, Test on Real) protocol. This involves a model trained exclusively on synthetic data and evaluated on real data. The difference from the model trained on real data (TRTR) measures the predictive utility of the synthetic data [24]. Jordon et al. [25] confirm that TSTR is currently the gold standard in this field, as it evaluates the preservation of relational structures between variables, rather than simply marginal distributions. Qian et al. [26] additionally propose the Total Variation Distance (TVD) within the Synthcity framework. They suggest calculating it variable by variable as the absolute difference between the proportions observed in the real and synthetic data. It constitutes a non-parametric metric directly applicable to binary variables. The threshold of 0.05 used as the limit for a negligible difference is consistent with the conventional threshold for statistical significance. Specifically in the educational context, a recent benchmark [27] comparing SMOTE and several deep generative models on 10,000 students shows that resampling methods achieve a TSTR of 0.997, confirming their predictive utility on tabular data. A TSTR score of 0.924, obtained on a dataset twenty times smaller and with a higher proportion of constrained binary variables, is consistent with this benchmark.

F. Handling Missing Values and Variable Selection

Missing value handling must precede any modeling of field data. Batista and Monard [28] have empirically demonstrated that KNN imputation produces statistically superior results compared to methods based on the mean or mode, as it leverages local relationships between variables to estimate missing values. Li et al. confirm that including the target variable in the search for neighbors improves the quality of imputation by leveraging available class information, a practice adopted in this pipeline. Emmanuel et al. [29] justify the exclusion of variables with more than 30% missing values, a threshold beyond which imputation introduces more noise than useful information.

Variable selection is a critical issue when the number of variables far exceeds the number of observations. Guyon and Elisseeff [30] have shown that combining several complementary methods reduces the risk of eliminating important variables due to the lack of power of a single method, justifying the three-method voting system with a unanimity requirement used in a context of n = 99 observations of the dataset [31]. Multicollinearity checking completes this step by considering the work of Dormann et al. [32] and Hair et al. They recommend a threshold of $|r| > 0.70$ as an exclusion criterion, above which two variables share more than 49% of their variance, making the regression coefficients unstable and their interpretation unreliable.

G. Evaluation Metrics and Classification Threshold

In the presence of imbalanced classes, the choice of metrics is a methodological decision. Japkowicz and Shah [33] demonstrate that a classifier that systematically predicts the majority class would achieve a high accuracy rate in the dataset without any discriminative value, rendering this metric inappropriate. The F1-score, the harmonic mean of precision and recall, is the primary recommended metric because it penalizes both types of errors equally. The AUC [34] provides an evaluation of overall discriminative ability independent of the threshold, which is particularly stable in cross-validation on small folds.

Determining the optimal classification threshold is distinct from model selection. The Youden index [35] simultaneously maximizes both detection rates under the assumption of symmetric costs between false positives and false negatives. In the considered career guidance context, this assumption is methodologically justified, since mistakenly directing a Developer toward Design and mistakenly directing a Designer toward Development represent equally costly failures for the candidate.

H. Similarity Measure for Mixed Variables and Person-Environment Fit Theory

Measuring similarity between profiles with continuous and binary variables requires an appropriate metric. Gower [36] proposed a general coefficient calculated as the average of individual similarities normalized according to the type of each variable, natively handling missing data and mixed types. D'Orazio [37] identifies a limitation of the standard version, noting that continuous variables with a wide interquartile range

tend to dominate the overall calculation. He proposes normalizing by the IQR rather than by the total range, thereby mitigating the impact of outliers and balancing the contribution of both types of variables. This distance is used in a Learn-Orient as a profile recognition filter before any prediction.

This filter is theoretically grounded in the concept of person–environment fit, formalized by Holland [38], which posits that professional satisfaction and success depend on the congruence between an individual's characteristics and the demands of their environment. This framework underpins the hypothesis that variables such as creativity scores or prior technical skills constitute stable indicators of natural fit for a particular field. Jaoui and Hilmi [39], in the French-speaking context of career transition, confirm that these fit factors remain robust determinants of career change decisions, regardless of formally acquired skills.

I. Positioning of the Present Study

The literature review presented in this chapter highlights four complementary gaps.

To begin with, the absence of a system for predicting and recommending career paths in the context of adult learning and career transition, particularly toward digital professions.

Second, existing work on career prediction using ML operates almost exclusively in contexts with large datasets containing standardized academic variables. No study offers a methodological solution for the simultaneous occurrence of a small sample size, predominantly binary variables with logical constraints, and imbalanced classes, a configuration that is characteristic of career transition data, particularly in an African context where data scarcity is also a common phenomenon.

Furthermore, while SMOTE has been validated for data augmentation on continuous educational data, its application to constrained binary variables remains unaddressed in the literature. The EDU-CDA method proposed in this study directly fills this gap by combining conditional SMOTE for continuous variables and conditional Bernoulli sampling for binary variables.

Finally, there is a lack of a component for detecting atypical profiles in existing guidance systems. No existing guidance system distinguishes between candidates whose profiles are covered by the model and those for whom the prediction is unreliable. The integration of the Gower distance normalized by the IQR as a pre-prediction filter constitutes an operational contribution that addresses this limitation.

These four gaps collectively define and establish the scope of this study, whose methodology is presented in the following chapter.

III. METHODOLOGY

A. Sampling and Data Collection Strategy

The target population of this study, individuals seeking career transition into digital professions, is not tracked by any official registry in Benin. No institution is formally dedicated to digital career transition in the strict sense. The phenomenon manifests itself in a diffuse manner across various training institutions (private digital schools or universities, vocational training centers, and integration programs) where some learners change careers or fields of study without this process being institutionally identified as such. The sampling universe is therefore, by nature, not delimited a priori, rendering any probabilistic approach inapplicable in the absence of an exhaustive sampling frame.

Considering this structural constraint, a non-probabilistic network sampling strategy, commonly referred to as snowball sampling, was adopted [40]. This method is recognized as appropriate when the target population is difficult to reach, unregistered, or distributed across multiple structures without administrative centralization [41], [42]. It involves identifying an initial group of participants who meet the inclusion criteria, then asking them to identify other potential candidates within their networks.

In the case of this study, the process began with direct contact with several institutions offering training in digital professions in Cotonou and surrounding cities, identified as likely to accept learners seeking career transition. A 47-item Google Forms questionnaire was distributed to these institutions and then shared by the initial respondents with their peers.

Two methodological precautions were taken to limit the biases inherent in this approach. At the outset, the deliberate diversification of entry points into the network. Several distinct institutions were officially contacted simultaneously to reduce the risk of overrepresentation of a particular subgroup linked to a single institution or a single social network. In addition, the anonymous and self-administered nature of the form limits social desirability bias in responses to questions about motivations and self-reported skills. Completing the form was voluntary, with no pressure from the administration. Respondents were not under any influence.

B. Data Collected and Sample Characteristics

This approach yielded 131 responses, of which 99 were retained after applying eligibility criteria that excluded programs with too few students for defensible statistical modeling (Audiovisual Specialist: 11 obs; Digital Project Manager: 8 obs; Data Analyst: 1 obs) and programs with ambiguous or incomplete titles. The primary inclusion criterion is the presence of an identifiable career break, where the candidate must have worked in or prepared for a career in a field distinct from digital professions before enrolling in their current program. This criterion operationalizes the concept of career transition regardless of the institution attended. The final dataset includes 61 candidates (61.6%) in the Graphic Design track and 38 (38.4%) in the Application Developer track, representing a 1.6:1 imbalance ratio. Table I presents the main characteristics of the sample.

C. Methodological Pipeline

The methodology is organized into seven sequential, interdependent steps, summarized in Table II. The dataset is divided into a training set (80%, 79 obs) and a test set (20%, 20 obs) via random sampling with a fixed seed. The test set is strictly isolated from the training data. It is not used in either variable selection or the construction of the EDU-CDA synthetic data, ensuring an unbiased evaluation.

TABLE I. SAMPLE CHARACTERISTICS BY TRACK (N = 99)

Characteristic	Graphic Designer (n=61)	Application Developer (n=38)
Average Age (years)	23.9 (σ=4.6)	24.6 (σ=5.4)
Male	67.2%	78.9%
From Benin	95.1%	97.4%
Family funding	70.5%	57.9%
Institutional funding	6.6%	18.4%
Personal funding	23.0%	23.7%
High school diploma or higher	34.4%	47.4%
Previous experience in the digital field	29.5%	50.0%
Availability > 72 hours/week	14.8%	26.3%
Average work experience	28.4 months	38.2 months

TABLE II. EDU-CDA METHODOLOGICAL PIPELINE

Step	Title	Input	Output
P1	Preprocessing & normalization	131 observations, 49 raw variables	99 obs, clean variables
P2	Variable coding	Textual/categorical variables	76 numerical variables
P3	Feature engineering	76 variables	80 variables (+ 4 derived variables)
P4	Variable selection (voting among 3 methods)	79 candidate variables	16 final variables
P5	Baseline modeling (real data)	79 real training observations	Baseline performance
P6	EDU-CDA increase	79 real training observations	1,600 balanced observations
P7	Final modeling + Evaluation	1,600 augmented observations	Evaluation of 20 real test observations

D. Preprocessing, Encoding, and Feature Engineering

Preprocessing begins with converting ambiguous textual responses (none, nothing, n/a, none, nr, nc, etc.) into missing values. The variable moy_bepc, which has an N/A rate of 35.4%, is excluded in accordance with the 30% threshold recommended by Emmanuel et al. [29]. The four remaining numerical variables (moy_bac, test_creativity, note_dipl, experience_month) are imputed using KNN (k = 5) with the target variable included in the distance calculation, which leverages the available class information [28].

The encoding transforms the textual variables using four complementary approaches: 1) simple binary encoding (gender, internet access, Beninese origin); 2) one-hot encoding with a priori categorical grouping for five variables, city (4 geographic zones), education level (Bac_Less / Bac_More), funding (4 types), field (3 categories), sector (3 categories); 3) multi-label encoding using regular expressions for multiple-choice variables (programming languages, design tools, office and cross-functional skills); (iv) lexical field detection for free-text responses (motivation, objective).

Four derived variables are constructed to synthesize the information scattered across individual binary columns with small sample sizes, improving the statistical power of the selection tests applied in the next step [30]:

$$\text{profile_dev} = \Sigma\ \text{programming_languages} + \text{html_level} \quad (1)$$

$$\text{designer_profile} = \Sigma\ \text{design_tools} + \text{photoshop_level} \quad (2)$$

$$\text{nb_trans} = \Sigma\ \text{cross-functional_skills} \quad (3)$$

$$\text{nb_office} = \Sigma\ \text{office_tools} \quad (4)$$

E. Selection of Variables by Unanimous Vote Using Three Methods

The selection is based on 79 variables available after feature engineering, using a voting system with three complementary methods [30]. A variable is retained only if it simultaneously satisfies all three criteria (score 3/3), which reduces the risk of spurious correlation in a sample of 99 observations [31]:

One-vs-all correlation: $r_j = \left|cor\left(X_j, Y_k\right)\right| > 0.10$, where Y_k is the binary indicator for sector k.

Random Forest Importance (MDA): 500 trees, average decrease in accuracy when permuting values > 0.

Statistical tests: Kruskal-Wallis (continuous variables); Fisher's exact test (binary variables, small sample sizes); $p < 0.05$.

Checking for multicollinearity at a threshold of $|r| > 0.70$ [33] led to the exclusion of 8 redundant variables as presented in Table III. The most informative variables within each set of redundant variables were retained.

TABLE III. EXCLUDED REDUNDANT PAIRS (|R| > 0.70)

Excluded variable	Retained variable	Observed |r|
high_school_level	high_school_level	1.000
html_level	dev_profile	0.928
Photoshop level	designer_profile	0.925
trans_proact	nb_trans	0.780
lang_html	dev_profile	0.775
illustrator_tool	designer_profile	0.765
no_trans	nb_trans	0.762
no_lang	dev_profile	0.752

The 16 final variables selected are presented in Table IV.

TABLE IV. 16 FINAL VARIABLES SELECTED (STEP P4)

Variable	Type	Category / Description
Age	Continuous	Socio-demographic
work_style	Continuous	Motivational (preference for solo vs. group)
logic_test	Continuous	Skill (logic test score)
creativity_test	Continue	Skill (creativity test score)
dev_profile	Continue	Derived — Σ languages + HTML level
designer_profile	Continue	Derived — Σ design tools + photoshop_level
nb_trans	Continue	Derivative — Σ cross-disciplinary skills

post-high-school level	Binary	Level of education ≥ High school diploma
end_of_family	Binary	Family funding
end_institutional	Binary	Institutional funding (scholarship)
time-intensive	Binary	Availability > 72 hours/week
Digital_Tech	Binary	Previous field: digital/tech
dom_other	Binary	Previous field: other
lang_java	Binary	Prior knowledge of Java
lang_js	Binary	Prior knowledge of JavaScript
trans_redac	Binary	Declared writing skills

F. EDU-CDA Method – C2 Contribution

The complete augmentation process is detailed in Fig. 1.

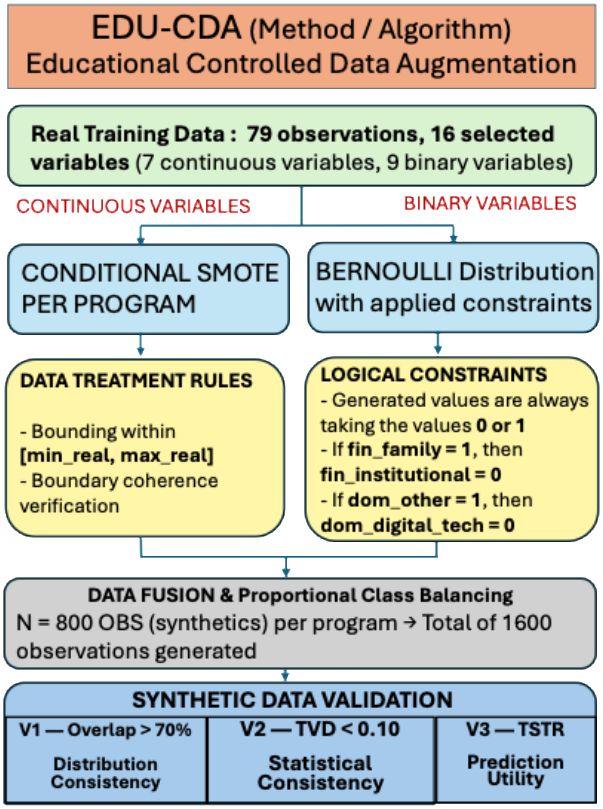


Fig. 1. EDU-CDA process.

EDU-CDA (Educational Controlled Data Augmentation) is one of this study's main contributions, developed to simultaneously address the dataset's three constraints: low volume (79 training observations), predominantly binary variables with logical constraints (9 out of 16), and imbalanced classes (1.6:1 ratio). It combines two generation mechanisms applied separately for each track k ∈ {Designer, Developer}, avoiding the generation of hybrid profiles at the boundary between the two classes.

1) Class-conditional SMOTE (continuous variables): For each track k, synthetic observations are generated by linear interpolation between a real observation x_i and one of its K nearest neighbors x_{nn} within the same class [10]:

$$\mathbf{x_{synth}} = x_i + \lambda \cdot (x_{nn} - \mathbf{x_i}),\ \lambda \sim U(0, 1) \quad (5)$$

The parameter K is adjusted: $K = \min(3, n_k - 1)$, avoiding outliers in classes with small sample sizes [11]. Two post-processing steps are systematically applied: 1) clamping values within the observed bounds $[x_{min}, x_{max}]$; 2) rounding to the nearest integer for count variables (profil_dev, profil_designer, nb_trans, test_logic, test_creativity, work_style).

2) Conditional Bernoulli per class (binary variables): For each binary variable x_j and each track k, the probability of occurrence is estimated from the actual training data:

$$\hat{p}_{jk} = \frac{1}{n_k} \cdot \sum_{i:y_i=k} x_{ij} \quad (6)$$

Each synthetic value is generated by a Bernoulli draw:

$$x_{ij}^{synth} \sim \text{Bernoulli}(\hat{p}_{jk}) \quad (7)$$

This approach guarantees values strictly ∈ {0, 1}, preserving the binary nature of the variables. It is preferred over SMOTE for binary variables whose interpolation would produce decimal values that are semantically unacceptable in the context of self-reported skills.

3) Consistency rules and augmentation strategy: Two business consistency rules correct logically impossible combinations: 1) if fin_family = 1, then fin_institutional = 0; 2) if dom_other = 1, then dom_digital_tech = 0.

Two augmentation strategies were used to determine the optimal strategy. The balanced strategy generates N synthetic observations per class regardless of the actual volume, resulting in approximately equal sample sizes across fields. The proportional strategy, on the other hand, generates $(N - n_k)$ synthetic observations per track, where n_k denotes the number of actual observations in class k, thereby ensuring exactly N observations per class in the final augmented dataset.

G. Validation of Synthetic Data

The quality of the synthetic data is evaluated according to three independent criteria, covering the distributional fidelity of continuous variables, the statistical fidelity of binary variables, and the predictive utility of the synthetic dataset.

1) Distributional fidelity (continuous variables): For each continuous variable, the overlap between the distributions of the real and synthetic data is calculated by intersecting the normalized histograms:

$$Overlap_j = \frac{\Sigma_b \left(h_j^{real}(b), h_j^{synth}(b)\right)}{\Sigma_b\, h_j^{real}(b)} \quad (8)$$

where, hj (b) denotes the density of variable j in bin b. An overlap greater than 70% is used as the acceptability criterion.

2) Statistical fidelity (binary variables): For each binary variable, the Total Variation Distance (TVD) between the real and synthetic proportions is calculated, Qian et al. [26]:

$$TVD_j = \left|\widehat{p_j^{réel}} - \widehat{p_j^{synth}}\right| \quad (9)$$

A TVD of less than 0.05 is considered negligible, by analogy with the conventional significance threshold α=0.05. A TVD of less than 0.10 is considered acceptable.

3) Predictive Utility (TSTR): An LDA model is trained exclusively on synthetic data and evaluated on real training data (TSTR), in accordance with the protocol by Esteban et al. [23]. Its accuracy performance is compared to that of an LDA model trained on real data only (TRTR). The choice of LDA as the reference classifier for this protocol is motivated by its stability across small samples. A difference of less than 0.15 is considered acceptable [43]:

$$\Delta_{TSTR} = |AccTRTR - AccTSTR| < 0{,}15 \quad (10)$$

H. Learn-Orient System - Contribution C1

The Learn-Orient system combines two sequential steps to produce a major recommendation with a graded confidence level. Fig. 2 illustrates the complete workflow.

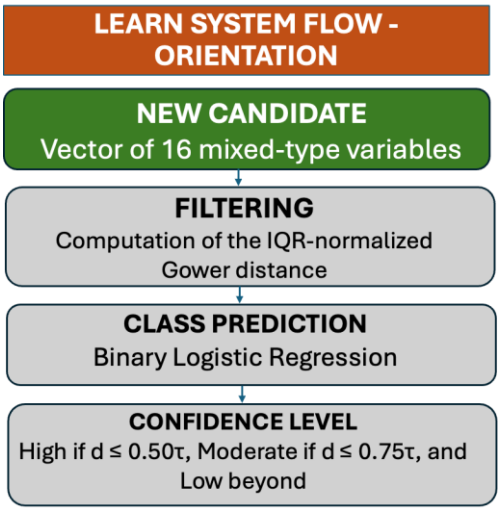


Fig. 2. Workflow of the learn-orient guidance system.

1) Profile recognition filter: The Gower distance normalized by the IQR [37] is calculated between the new candidate's profile and each of the 1,600 reference observations. For continuous variables, partial similarity is normalized by the IQR (rather than the range) to mitigate the impact of outliers and balance the contribution of both types of variables. For binary variables, the asymmetric Jaccard coefficient is used so that a double absence does not contribute to the similarity. The average distance to the k = 5 nearest neighbors is calculated as follows:

$$\underline{d}_G ower(x_n ew) = (1/5) \cdot \Sigma_{j=1}^{5} d_G ower\left(x_n ew, x_{(j)}\right) \quad (11)$$

A threshold τ is defined as the 95th percentile of the distribution of mutual distances between reference observations. If $\bar{d} > \tau$, the profile is declared unrecognized, and the system recommends considering other digital career paths besides designer and developer, thereby signaling its own limitations for atypical profiles (RQ4).

2) Binary logistic regression and confidence levels: If the profile is recognized, binary logistic regression estimates the membership probabilities:

$$P(x) = 1/\left(1 + exp\left(-(\beta_0 + \beta^T x)\right)\right) \quad (12)$$

$$P(Designer \mid x) = 1 - P(Developer \mid x) \quad (13)$$

The β coefficients are estimated using the augmented dataset (N = 1,600) and interpreted as odds ratios $OR_j = \exp(\beta_j)$ [34]. An odds ratio greater than 1 indicates a positive association with the Application Developer track; an odds ratio less than 1 indicates a positive association with the Graphic Designer track. At the same time, confidence intervals are evaluated to identify the variables significant for predicting the targeted tracks. If the interval $[CI_{lower}, CI_{upper}]$ contains 1, the variable is considered insignificant.

The optimal threshold is determined by the Youden index [35]: J = Sensitivity + Specificity − 1. Three confidence levels are assigned based on $\bar{d} / \tau$:

- High (≤ 0.50), reliable automatic prediction;
- Moderate (≤ 0.75), recommendation for further evaluation;
- Low (> 0.75 but ≤ 1.00), maintenance strongly recommended [10].

I. Algorithms

Seven binary classification algorithms are compared in two phases: a baseline phase on real-world data (P5) and a post-augmentation phase on EDU-CDA data (P7). The evaluated algorithms are binary logistic regression (LR), logistic regression with class weighting (LR + Weighting), radial basis support vector machine (SVM), random forest (RF), linear discriminant analysis (LDA), naive Bayes classifier, and decision tree (CART).

The dataset is divided into a training set (80%, or 79 observations) and a test set (20%, or 20 observations) via simple random sampling with a fixed seed (set.seed(42)). The test set is used only once, for the final evaluation, and is not involved at any stage in the construction of synthetic data or in variable selection.

J. Evaluation Metrics

1) Supplementary metrics: Five complementary metrics are calculated for each model. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, with the positive class being the Application Developer track.

Accuracy: Overall proportion of correct predictions, provided for informational purposes only. In the presence of imbalanced classes, a classifier that systematically predicts the majority class can achieve high accuracy without discriminative value [33].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Sensitivity and Specificity: Sensitivity measures the proportion of correctly identified Application Developers; specificity measures the proportion of correctly identified Designers:

$$Sensitivité = \frac{TP}{TP+FN} \qquad Spécificité = \frac{TN}{TN+FP} \tag{15}$$

F1-score: Primary model selection metric, defined as the harmonic mean of precision and recall, suitable for imbalanced problems [33]:

$$F1 = 2 \cdot \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{16}$$

AUC: Area under the ROC curve, measuring overall discriminatory power independently of the threshold [34]. Primary metric in cross-validation due to its stability on small folds. Train-Test Gap: An indicator of overfitting [44]:

$$\Delta_{overfitting} = F1_{train} - F1_{test} \tag{17}$$

A significant discrepancy between training and test performance indicates that the model has memorized the data rather than learned generalizable patterns.

2) Cross-validation: Three complementary 5-fold cross-validations are conducted using the caret package.

CV1 — On real data. A 5-fold cross-validation is performed on the 79 real training observations to estimate the performance of each algorithm more robustly than a single split, whose random composition on a small sample can produce unstable estimates. AUC is chosen as the evaluation metric due to its stability on small folds.

CV2 — On augmented data. A 5-fold cross-validation is performed on the 1,600 augmented observations to assess the stability of the final model's performance on the synthetic data space. Both cross-validations are implemented using the `trainControl`.

CV3 — On real data after augmentation. A third cross-validation complements CV1 and CV2 to quantify the reliability of the performance estimates and to test the significance of the differences between models. CV3 is a stratified, repeated cross-validation (5 folds, 10 repetitions, i.e., 50 evaluations per model). Unlike CV1, which trains on the real data alone, and CV2, which trains and tests on the augmented data, EDU-CDA augmentation is applied only to the training partition of each fold, the test fold remaining composed exclusively of real observations. The distribution of the 50 F1 scores obtained per model is used to estimate a 95% confidence interval. Finally, the differences between models are tested using the exact McNemar test applied to the out-of-fold predictions: for each pair of models, the test examines only the observations on which the two models disagree and checks whether these disagreements lean significantly in favour of one or the other.

K. Tools and Technical Environment

The analyses were performed using R (version 4.x) and Python 3.x via Google Colaboratory. In R, the main packages used were: caret (modeling and cross-validation), VIM (KNN imputation), ranger (random forests), glm (logistic regression), StatMatch (Gower's distance), and pROC (ROC curves and AUC). In Python, the scikit-learn, imbalanced-learn, pandas, and numpy libraries were used for preprocessing and algorithm comparison. Reproducibility is ensured by setting the random seeds (set.seed(42) in R, random_state=42 in Python).

IV. Results

A. Variable Selection (RQ1)

Out of 79 candidate variables, the unanimous vote across three methods selected 24 variables. As shown in Fig. 3, after checking for multicollinearity, 16 final variables were selected (Table IV). The correlation heatmap confirms the overall independence of the 16 final variables. The strong correlations observed are expected and consistent: fin_famille and fin_institutionnel show a negative correlation (mutually exclusive sources). test_logique and test_creativite show a moderate positive correlation. profil_dev and profil_designer do not have a strong correlation between them, confirming their ability to distinguish the two tracks independently. The absence of strong correlations among the other variables validates the quality of the P4 selection.

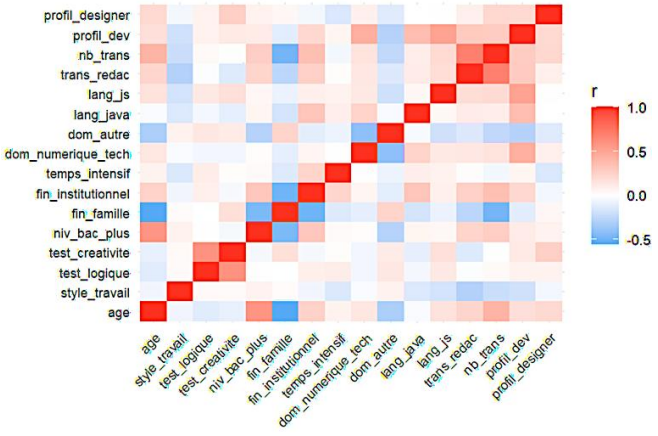


Fig. 3. Heatmap of 16 selected variables.

Profil_dev has the highest correlation with the target (r = 0.427) and the most pronounced difference in means (Designer: 1.13 vs. Developer: 3.13). profil_designer shows the inverse relationship (r = −0.291). These derived variables effectively synthesize the information scattered across individual binary columns with small sample sizes.

B. Validation of EDU-CDA Synthetic Data (RQ3)

The three criteria for tripartite validation are fully satisfied, providing a positive answer to RQ3 as shown in Table V.

TABLE V. Tripartite Validation of EDU-CDA Synthetic Data

Criterion	Indicator	Threshold	Result	RQ3
Distribution accuracy	Histogram overlap (7 continuous variables)	> 70%	82.9–90.8%	Yes
Statistical accuracy	TVD proportions (9 binary variables)	< 0.10	Max. = 0.061	Yes
Predictive utility	TSTR: ΔAccuracy LDA (synthetic vs. real)	Δ < 0.15	Δ = 0.025	Yes

C. Comparative Performance of the Models (RQ1, RQ2)

Table VI and Fig. 4 present the performance of the seven algorithms on the real test set. In the baseline phase (P5), LDA is the model with the best F1 score (F1 = 0.889, AUC = 0.940, ΔTrain = 0.060). Binary LR and RF achieve a perfect training score (1.000), indicating severe overfitting due to near-linear separability in a 16-variable space across 79 observations [38]. After EDU-CDA augmentation (P7), Binary LR is the only model to achieve F1 = 0.900 while maintaining the lowest ΔTrain among the competitive models (0.080), positively addressing RQ2. RF retains persistent overfitting (ΔTrain = 0.222). The Binary LR is selected as the production model.

TABLE VI. PERFORMANCE ON REAL TEST SET (20 OBS) ON P5 AND P7

Model	Acc.	F1	Sens.	Spec.	AUC	Ph	ΔTrain
LDA ★	0.90	**0.889**	0.80	1.00	0.940	P5	0.060
SVM	0.85	0.842	0.80	0.90	0.920	P5	0.133
LR Binary	0.85	0.824	0.70	1.00	0.860	P5	0.176
RF	0.85	0.842	0.80	0.90	0.920	P5	0.158
LR Bin. ★	0.90	**0.900**	0.90	0.90	0.950	P7	0.080
SVM	0.85	0.857	0.90	0.80	0.960	P7	0.136
LDA	0.85	0.842	0.80	0.90	0.950	P7	0.132
RF	0.80	0.778	0.70	0.90	0.930	P7	0.222
CART	0.85	0.842	0.80	0.90	0.855	P7	0.095
Naive Bayes	0.75	0.706	0.60	0.90	0.875	P7	0.214

[a] ★ = Best model of the phase (Ph). ΔTrain = F1_train − F1_test. Sens. = Sensitivity. Spec. = Specificity

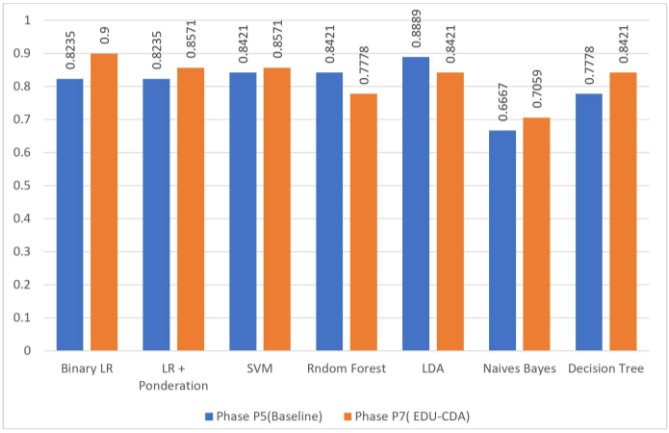


Fig. 4. Comparative F1 performance — Phases P5 (Baseline) and P7 (Post-EDU-CDA) on the real test set – before and after augmentation.

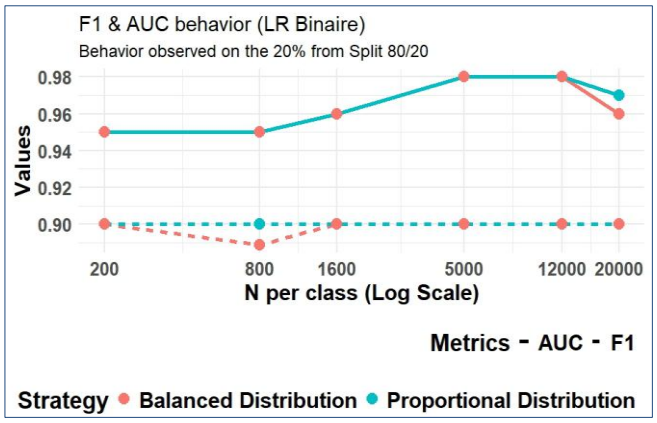


Fig. 5. Analysis of sensitivity to the volume of augmentation.

Three observations emerge from this sensitivity analysis. Initially, F1 plateaus at N = 200 for the proportional strategy (F1 = 0.900 across all configurations) as observed in Fig. 5. The model reaches its maximum binary discrimination capacity very quickly. Next, the AUC follows an upward trajectory up to N = 5,000–12,000 (AUC = 0.98), reflecting a gradual improvement in probability calibration without changing the binary decisions.

In addition, the balanced strategy exhibits a local degradation at N = 800 (F1 = 0.889, Sensitivity = 0.80), suggesting instability in the generation method at this volume level. The proportional strategy is therefore preferred.

D. Cross-Validation and Sensitivity Analysis (RQ2)

5-fold cross-validation on real data (CV1) confirms the hierarchy: LDA is the best model (AUC = 0.944, σ = 0.029), while Binary LR has the lowest AUC (0.810, σ = 0.086). Cross-validation on augmented data (CV2) yields remarkable stability, as shown in Table VII. The increase in AUC from 0.810 to 0.995 provides empirical evidence of the contribution of EDU-CDA.

TABLE VII. CROSS-VALIDATION ON AUGMENTED DATA (BINARY LR, N=800)

Fold	AUC	Sensitivity	Specificity
Fold 1	0.9926	0.9688	0.9688
Fold 2	0.9967	0.9875	0.9813
Fold 3	0.9964	0.9750	0.9750
Fold 4	0.9985	0.9750	0.9875
Fold 5	0.9929	0.9438	0.9938
Average ★	**0.9954**	0.9700	0.9813
Standard deviation	0.0026	0.0162	0.0099

The influence of the volume of synthetic data on the performance of the binary LR model was systematically studied across six levels of N (200 to 20,000 observations per class) and two generation strategies. This evaluation, presented in Table VIII, was conducted exclusively on the 20 real observations in the test set to ensure the external validity of the results.

TABLE VIII. CROSS-VALIDATION FOR EACH DISTRIBUTION AND N

Strategy	N/class	Total obs	AUC	SD
Proportional	200	400	0.9815	0.0153
Proportional	800	1,600	0.9954	0.0026
Proportional	1,600	3,200	0.9940	0.0034
Proportional	5,000	10,000	0.9946	0.0004
Proportional	20,000	40,000	0.9945	0.0008
Balanced	200	479	0.9870	0.0124
Balanced	800	1,679	0.9948	0.0034
Balanced	5,000	10,079	0.9951	0.0009
Balanced	20,000	40,079	0.9944	0.0014

The cross-validation stability analysis based on the increase in sample size shows that N = 200 exhibits significant instability for both strategies, with high standard deviations (0.0153 and 0.0124, respectively), indicating that the model has not yet stabilized at this level. Starting at N = 800, the proportional strategy simultaneously achieves the highest AUC (0.9954) and the lowest SD (0.0026).

N = 800 with the proportional strategy emerges as the optimal configuration: it is the only point where the model combines a maximum AUC, a minimum SD, and stability demonstrated fold by fold, without requiring an excessive volume of data.

As observed in Fig. 6, the best-ranked model is LDA (F1 = 0.831), closely followed by Binary LR (F1 = 0.805). The confidence intervals of the five best models overlap substantially, ruling out any reliable ranking.

```
=== F1 Mean + IC 95% ===
> print(ci, row.names = FALSE)
                Modele    F1 IC_inf IC_sup
                   LDA 0.831  0.726  0.935
             LR Binaire 0.805  0.690  0.921
 LR + Pondération 0.801  0.686  0.916
    Random Forest 0.794  0.666  0.921
                   SVM 0.790  0.672  0.908
    Decision Tree 0.720  0.593  0.847
      Naive Bayes 0.705  0.555  0.855
```

Fig. 6. Confidence interval (screenshot on terminal).

Two consequences follow. CV3 shows that EDU-CDA raises Binary LR, the most penalized model on real data alone (CV1: AUC = 0.810), to statistical parity with the best baseline model (LDA), corroborating the contribution of EDU-CDA (H2) while qualifying the idea of a strict ranking. Fig. 7 shows the results of applying the McNemar test to the out-of-fold predictions.

```
=== McNemar – LDA vs OTHER MODELS ===
> print(do.call(rbind, lapply(setdiff(modeles, tete),
+                             function(m) mcnemar_pair(tete, m))), row.names = FALSE)
             Comparison A_best B_best      p Significant
      LDA vs LR Binaire      3      3 1.0000         Non
LDA vs LR + Pondération      3      3 1.0000         Non
             LDA vs SVM      7      2 0.1797         Non
   LDA vs Random Forest      9      4 0.2668         Non
     LDA vs Naive Bayes      9      1 0.0215         Oui
   LDA vs Decision Tree     15      6 0.0784         Non
```

Fig. 7. McNemar test applied to the out-of-fold predictions.

Binary LR is selected as the final model for two reasons, independent of raw performance. To begin with, it makes no assumption about the distribution of the predictors, whereas LDA assumes multivariate normality within each group — an assumption ill-suited to a predominantly binary dataset (9 variables out of 16). Second, it provides odds ratios together with their confidence intervals, offering a direct and auditable interpretation of the orientation factors that LDA does not allow in such a readable form.

E. Final Model - Optimal Threshold and Odds Ratios (RQ1)

The Youden index identifies the threshold 0.6061 (Sensitivity = Specificity = 0.90) as optimal. Table IX and Fig. 8 present the odds ratios of the final model estimated on the augmented data (N = 800, 95% CI).

TABLE IX. ODDS RATIOS OF THE FINAL BINARY LR MODEL (N=800, 95% CI)

Cariable	OR	IC Inf.	IC Sup.
(Intercept)	52,729	1,887	1698,036
age	1,445	1,303	1,629
style_travail	0,433	0,294	0,632
test_logique	1,648	1,181	2,362
test_creativite	0,107	0,058	0,175
niv_bac_plus	2,521	1,136	5,826
fin_famille	0,477	0,213	1,036
fin_institutionnel	32,399	3,021	657,726
temps_intensif	3,199	1,353	7,979
dom_numerique_tech	3,402	1,546	7,759
dom_autre	0,898	0,329	2,464
lang_java	29,913	3,643	498,063
lang_js	1,123	0,266	6,189
trans_redac	2,552	1,168	5,791
nb_trans	1,402	1,006	1,983
profil_dev	5,628	4,066	8,346
profil_designer	0,178	0,122	0,244

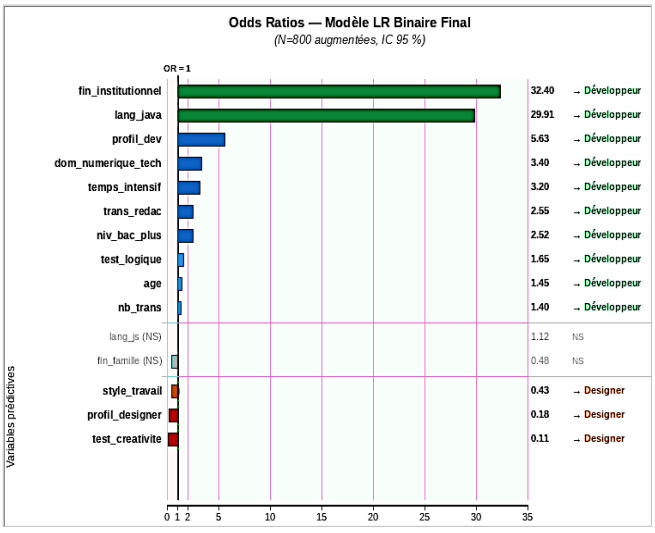


Fig. 8. Odds ratios of the final binary LR model.

F. Class-Wise Analysis and Confusion Matrix (RQ1)

Table X presents the final model's performance by class on the real test set (20 observations). Both classes exhibit symmetrical performance (Precision = Recall = F1 = 0.90), reflecting the quality of the EDU-CDA balancing and the effectiveness of the chosen Youden threshold.

TABLE X. PERFORMANCE BY CLASS VIA BINARY LR AND POST-EDU-CDA (P7) ON THE REAL TEST SET

Class	Precision	Recall	F1-score	Support
Graphic Designer	0.90	0.90	0.90	10
Application Developer	0.90	0.90	0.90	10
Macro average	**0.90**	**0.90**	**0.90**	20
Overall accuracy	—	—	**0.90**	20

The Binary LR after EDU-CDA is the only model to exhibit a perfectly symmetrical error distribution (1 FN, 1 FP), confirming the sensitivity/specificity balance (0.90/0.90) observed in the aggregated metrics. Table XI presents this performance.

TABLE XI. PERFORMANCE BY CLASS BY MODEL

Model	VP	FN	FP	VN
LR Binary ★	9	1	1	9
SVM	9	1	2	8
LR + Weighting	9	1	2	8
LDA	8	2	1	9
Decision Tree	8	2	1	9
Random Forest	7	3	1	9
Naive Bayes	6	4	1	9

After boosting, Binary LR is the only model to achieve F1 = 0.900 on the test set while maintaining the lowest training-test gap among competitive models (0.080). Random Forest continues to exhibit persistent overfitting (gap = 0.222). Binary LR is selected as the production model. Table XII presents the confusion matrix.

TABLE XII. CONFUSION MATRIX BINARY LR POST-EDU-CDA (P7)

Model	F1 Training	F1 Test	Gap
Random Forest	1.000	0.778	0.222
SVM	0.993	0.857	0.136
LDA	0.974	0.842	0.132
Binary LR	0.980	0.900	0.080

G. Final Learn-Orient Classification Model (RQ4)

Applying the Youden index to the ROC curve of the final model yields two thresholds at J = 0.80. The graph in Fig. 9 illustrates the ROC curve of the final model. It has an AUC of 0.95, indicating excellent overall discriminatory power. It deviates significantly from the reference diagonal (AUC = 0.50 corresponding to a random classifier), confirming that the model effectively distinguishes between Designer and Developer profiles across all possible thresholds.

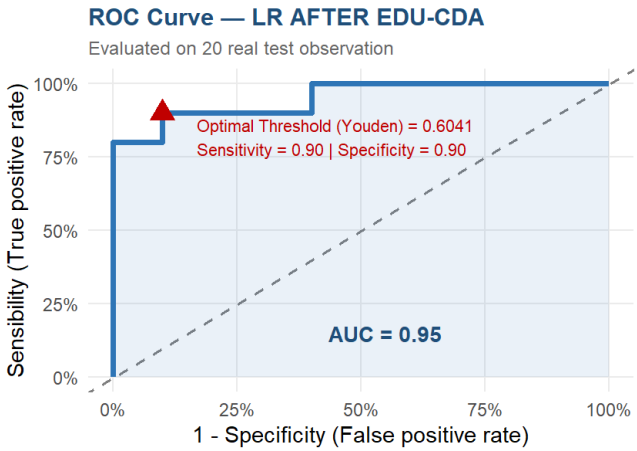


Fig. 9. ROC Curve of the final model after EDU-CDA applied.

The threshold of 0.6061 is selected as the optimal threshold. When two thresholds achieve the same Youden index, statistical convention recommends selecting the lower one, which in this case corresponds to a perfect symmetrical balance between sensitivity and specificity (0.90/0.90). Presented in Table XIII, this balance is particularly well-suited to the context of career guidance, where the two types of error, incorrectly directing a Developer toward Design and incorrectly directing a Designer toward Development, have comparable consequences for the candidate.

TABLE XIII. OPERATIONAL TESTS - LEARN-ORIENT SYSTEM ON 3 PROFILES

Profile Tested	$\bar{d}$ Gower	% threshold τ	Prediction	Confidence
Test 1 — Actual designer	0.092	15%	Designer 99.9%	High
Test 2 — Actual developer	0.175	28%	Developer 100.0%	High
Test 3 — Atypical profile	0.377	60%	Developer 100.0%	Moderate
Gower threshold $\tau = 0.629$	—	100%	—	—

The real profiles (Tests 1 and 2) have Gower distances less than 28% of the threshold $\tau = 0.629$, with clearly defined probabilities and High confidence. The artificial atypical profile (Test 3) correctly triggers a Moderate confidence level at 60% of the threshold, answering RQ4 in the affirmative.

V. DISCUSSION

A. Discriminability of Pre-Training Profiles (RQ1)

Learn-Orient achieves F1 = 0.900 and AUC = 0.950 on the real test set without standardized academic variables, answering RQ1 positively with performance well above the operational threshold. This result is even more notable as it is obtained on only 99 observations and surpasses the ceiling observed by Al-Dossari et al. [12] (70.47% with XGBoost on a dataset twenty times larger). It confirms that the performance of a guidance system does not depend simply on the volume of data, but on the relevance of the variables and the quality of the augmentation method.

Two counterintuitive predictors deserve attention. Institutional funding (OR = 32.40) outperforms all technical skills: this is not an intrinsic characteristic of the candidate but a characteristic of their context of access to training. This strong signal suggests that government scholarship programs structurally attract technical profiles, targeting fields perceived as strategic (software development). This result opens important perspectives for policies on funding retraining.

The creativity test score (OR = 0.107) can be interpreted considering Holland's Person-Environment Fit theory [38]: self-reported creativity is a stable indicator of fit with the Graphic Designer profile, regardless of formal technical skills. Recent studies in the French-speaking context of career transition [39] confirm that these person–environment fit factors remain robust determinants of career change decisions. Song et al. [17] similarly emphasize the importance of motivational and behavioral variables beyond academic characteristics alone.

B. Effectiveness of the EDU-CDA Method (RQ2)

The Binary Logistic Regression model test on the test set of 20 observations yields a F1 = 0.900. This represents a meaningful indicator of the validity of the EDU-CDA Method. It can be explained by a well-documented principle [10] which establishes that the simplest models, those most vulnerable to small samples, benefit the most from augmentation. Before EDU-CDA, quasi-linear separation on 79 observations produced a perfect training score (F1_train = 1.000). After augmentation to 1,600 observations, this phenomenon disappears (F1_train = 0.980), and the model generalizes correctly.

EDU-CDA addresses a specific shortcoming of standard SMOTE [11] : its inapplicability to constrained binary variables. The TSTR of 0.924, obtained on a dataset twenty times smaller and predominantly binary than that of Chinodakufa et al. [27] (TSTR = 0.997 on 10,000 continuous observations), falls within defensible values. The results of Anggrawan et al. [21] and Cu et al. [22] show similar stabilization patterns from equivalent augmentation volumes.

CV3 cross-validation applies the augmentation within each fold and tests on real data. It confirms that EDU-CDA moves Binary LR from the status of the most vulnerable model to statistical parity with the best baseline model (LDA): across the 99 real observations, Binary LR and LDA do not differ significantly in any case (McNemar test, p = 1.000). The realistic generalization performance of Binary LR stands at around F1 = 0.80 [0.69; 0.92], the value of 0.900 obtained on the single test set reflecting a favourable composition of the split. EDU-CDA, therefore, does improve Binary LR.

C. Fidelity of Synthetic Data and the Gower Filter (RQ3 and RQ4)

The three-part protocol fully validates RQ3. The only TVD value slightly exceeding 0.05 (end_family, TVD = 0.061) concerns the variable whose highest proportion (65.7% family funding) makes exact reproduction via conditional Bernoulli more sensitive to random fluctuations, an expected behavior with no practical impact on overall quality. The combination of these three dimensions into a single protocol had not been formalized in the literature on educational data.

Operational tests validate RQ4: the Gower filter correctly distinguishes well-covered profiles (distances < 28% of the threshold) from atypical profiles (60% of the threshold). The IQR normalization proposed by D'Orazio [37] is essential: without it, continuous variables with high interquartile ranges would dominate the calculation, mechanically reducing the contribution of binary variables.

D. Limitations

While the proposed model and methodology demonstrate good performance, some limitations are acknowledged:

L1- Representativeness: the sample obtained does not claim to be statistically representative of all individuals undergoing digital retraining in Benin. The validity of the study rests on the relevance and diversity of the collected profiles rather than on their representativeness in the probabilistic sense. The results and the developed system constitute a proof of concept, and a decision-support tool based on the patterns observed in this sample; generalizing these findings to other contexts will require external validation on new cohorts. CV3 cross-validation also shows that the five best models are statistically indistinguishable at this sample size. The F1 differences reported in phase P7 should therefore be read as indicative, and confirmation on a larger sample remains desirable.

L2 — Statistical power: With 20 observations in the test set, a misclassified observation accounts for 5 F1 points. Confirmation is therefore required on a larger sample.

L3 — Sample bias: overrepresentation of males (71.9%) and near-exclusivity of Beninese individuals (95.6%) amplified in the synthetic data.

L4 — Partial conditional independence: dependencies between binary variables not covered by the two consistency rules are not preserved by the Bernoulli component.

L5 — Self-reporting bias: technical skills are not objectively verified, constituting a source of unmeasurable noise.

L6 — Restriction to two fields: Audiovisual Specialist (n=11) and Digital Project Manager (n=8) remain outside the scope due to insufficient sample sizes.

L7 — Gower threshold calculated on synthetic data: the threshold τ characterizes the internal dispersion of the augmented space. External validation on new cohorts is necessary.

L8 — Lack of hyperparameter optimization: the algorithms are evaluated using their default parameters. Optimization could alter the performance hierarchy for the SVM and RF.

L9 — Discriminant variable selection criteria: Selecting variables based on three criteria could exclude important variables with a small number of observations.

VI. CONCLUSION

Career guidance toward digital professions in Sub-Saharan Africa remains, to date, a largely subjective, non-scalable, and poorly equipped process. This study builds on this concrete observation to propose a response that is both scientific and operational, grounded in real data from different programs and training centers or universities in Benin.

The first contribution of this study (Learn-Orient) demonstrates that it is possible to statistically distinguish between two career transition profiles based exclusively on pre-training characteristics: without grades, without formal assessment, and without structured interviews. With an F1 score of 0.900 and an AUC of 0.950 on real-world data, the system generates interpretable probabilistic recommendations, accompanied by a graded confidence level that explicitly signals its own limitations. The Gower filter ensures responsible use by identifying atypical profiles. This last point, beyond being a technical detail, serves as a genuine ethical safeguard. A system that knows when not to respond is more useful and more responsible than a system that always responds.

The second contribution (EDU-CDA) addresses a specific methodological challenge that the literature had not yet resolved: how to augment educational data characterized by constrained binary mixed variables, low volume, and imbalanced classes? The proposed solution is simple in its logic, rigorous in its implementation, and validated across three independent dimensions. It transformed logistic regression, the most vulnerable model on raw data, into the most stable model after augmentation, with an AUC rising from 0.810 to 0.995 in cross-validation. This result suggests that simple, interpretable models are not inherently inferior to complex models; they generally require more data to realize their full potential.

Beyond the metrics, the results yield operational insights that decision-makers can directly apply. The fact that institutional funding is the strongest predictor of the Developer profile (OR = 32.40), ahead of all technical skills, is not a statistical anomaly. This would suggest that training in this field is costly and that support in this area is likely the best way to provide high-quality, large-scale training for this profession. It also signals that funding policies are already, unintentionally, shaping the profiles of those entering certain fields. Similarly, the creativity score as the primary predictor of the Designer profile (OR = 0.107) confirms that cognitive abilities measured before training are robust indicators of career orientation, regardless of prior academic background.

In the context of using Learn-Orient, these results call for four concrete actions: institutionalize pre-admission profiling using the 16 identified variables; differentiate funding mechanisms according to the targeted fields of study; use the atypical profile signal as a trigger for enhanced human support; and actively diversify the pool of candidates to correct current demographic biases.

This study is a first step, not an endpoint. Extending the system to other fields, such as Audiovisual Specialist, Digital Project Manager, Data Analyst, and others, as soon as staffing allows, would broaden the system's coverage. Integrating a component that estimates the probability of success conditional on the field of study would transform Learn-Orient into a true recommendation tool rather than merely a profile similarity tool. External validation on new training cohorts, followed by deployment as a web application accessible to educational advisors, constitute the two immediate operational steps. Finally, the EDU-CDA method is designed to be transferable without major modification to similar programs, particularly in Africa. It is worth noting that it requires neither a large-scale data collection infrastructure nor expertise in deep generative models.

In a context where 230 million digital jobs are expected in sub-Saharan Africa by 2030 [5], properly guiding each candidate from the outset is not a methodological luxury. It is a direct lever for employability, equitable access, and the efficiency of resources invested in training.

Ethical Considerations

All data was collected with the informed consent of the respondents. The data is anonymized: no personally identifiable information is used in the analyses, and the results are presented exclusively in aggregated form. Learn-Orient is designed as a decision-support tool complementary to human assessment, not as an autonomous decision-making system. The educational advisor retains full responsibility for the final guidance. The profile recognition filter ensures that the system flags its own limitations for atypical profiles. In accordance with Rudin's [15] recommendations for systems with high human stakes, the odds ratios from the logistic regression allow the counselor to explain the reasons behind the generated recommendation to the candidate with the utmost transparency.

Acknowledgment

Our sincere thanks go to Firmin Monyevedo TOVODOUNNON, CEO of EIG Group, Issiakou SOULEYMANE, CEO of the School of Digital Professions (École des Métiers du Numérique - EMN) in Benin, all other training centers and schools that have shared their data with us, as well as all the candidates who agreed to participate in this study. This work is part of the research activities of the Computer Science Research and Decision Support Laboratory (L@RIAD), from the Institute of Mathematics and Physical Sciences (IMSP), an entity of the University of Abomey-Calavi (UAC), whose supervisory team we would like to thank.

References

[1] R. A. Marte Espinal et L. Fabián V, « Determinantes de la deserción universitaria: un estudio de caso en la República Dominicana », SIJIS, vol. 2, no 1, p. 255-268, mars 2021, doi: 10.51798/sijis.v2i1.76.

[2] G. Wydra-Somaggio, « Early termination of vocational training: dropout or stopout? », Empirical Res Voc Ed Train, vol. 13, no 1, p. 5, déc. 2021, doi: 10.1186/s40461-021-00109-z.

[3] International Labour Office, Global Employment Trends for Youth 2022: investing in transforming futures for young people. S.l.: INTL LABOUR OFFICE, 2022.

[4] OECD, Dir., OECD Digital Economy Outlook 2024 (Volume 2): Strengthening Connectivity, Innovation and Trust. dans OECD Digital Economy Outlook. Paris: OECD Publishing, 2024. doi: 10.1787/3adf705b-en.

[5] IFC - World Bank, Empowering Africa's youth: Bridging the digital skills gap. 2121 Pennsylvania Avenue, NW Washington, DC 20433 USA: World Bank, 2024.

[6] M. S. Macharia, « Reengineering mass career acquisition through technical vocational education training counseling in Kenya », IJRBS, vol. 8, no 6, p. 212-218, oct. 2019, doi: 10.20525/ijrbs.v8i6.533.

[7] F. Eicker, G. Haseloff, et B. Lennartz, Vocational Education and Training in Sub-Saharan Africa: Current Situation and Development. Bielefeld: W. Bertelsmann Verlag, 20170316.

[8] G. Adankon et P. Houngue, « How Can Intelligent Educational Technologies Address the Challenges of Professional Retraining in Africa: From Current State to Future Perspectives? », dans 2025 IEEE

Global Engineering Education Conference (EDUCON), London, United Kingdom: IEEE, avr. 2025, p. 1-8. doi: 10.1109/EDUCON62633.2025.11016638.

[9] F. Trujillo, M. Pozo, et G. Suntaxi, « Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction », J. Technol. Sci. Educ., vol. 15, no 1, p. 162, mars 2025, doi: 10.3926/jotse.3124.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer, « SMOTE: Synthetic Minority Over-sampling Technique », jair, vol. 16, p. 321-357, juin 2002, doi: 10.1613/jair.953.

[11] Haibo He et E. A. Garcia, « Learning from Imbalanced Data », IEEE Trans. Knowl. Data Eng., vol. 21, no 9, p. 1263-1284, sept. 2009, doi: 10.1109/TKDE.2008.239.

[12] H. Al-Dossari, F. A. Nughaymish, Z. Al-Qahtani, M. Alkahlifah, et A. Alqahtani, « A Machine Learning Approach to Career Path Choice for Information Technology Graduates », Eng. Technol. Appl. Sci. Res., vol. 10, no 6, p. 6589-6596, déc. 2020, doi: 10.48084/etasr.3821.

[13] P. Houngue, M. Hountondji, et T. Dagba, « An Effective Decision-Making Support for Student Academic Path Selection using Machine Learning », IJACSA, vol. 13, no 11, 2022, doi: 10.14569/IJACSA.2022.0131184.

[14] M. Wu, G. Subramaniam, D. Zhu, C. Li, H. Ding, et Y. Zhang, « Using Machine Learning-based Algorithms to Predict Academic Performance - A Systematic Literature Review », dans 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India: IEEE, févr. 2024, p. 1-8. doi: 10.1109/ICIPTM59628.2024.10563566.

[15] C. Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », Nat Mach Intell, vol. 1, no 5, p. 206-215, mai 2019, doi: 10.1038/s42256-019-0048-x.

[16] D. W. Hosmer, S. Lemeshow, et R. X. Sturdivant, Applied Logistic Regression, 1re éd. dans Wiley Series in Probability and Statistics. Wiley, 2013. doi: 10.1002/9781118548387.

[17] Q. C. Song, H. J. Shin, C. Tang, A. Hanna, et T. Behrend, « Investigating machine learning's capacity to enhance the prediction of career choices », Personnel Psychology, vol. 77, no 2, p. 295-319, juin 2024, doi: 10.1111/peps.12529.

[18] L. K. Smirani, H. A. Yamani, L. J. Menzli, et J. A. Boulahia, « Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths », Scientific Programming, vol. 2022, p. 1-15, avr. 2022, doi: 10.1155/2022/3805235.

[19] J.-E. Lee, A. Jindal, S. N. Patki, A. Gurung, R. Norum, et E. Ottmar, « A comparison of machine learning algorithms for predicting student performance in an online mathematics game », Interactive Learning Environments, vol. 32, no 9, p. 5302-5316, oct. 2024, doi: 10.1080/10494820.2023.2212726.

[20] F. E. Arévalo-Cordovilla et M. Peña, « Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors », Mathematics, vol. 12, no 20, p. 3272, oct. 2024, doi: 10.3390/math12203272.

[21] Information Technology Education Department, Bumigora University, Indonesia, A. Anggrawan, H. Hairani, et C. Satria, « Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE », IJIET, vol. 13, no 2, p. 289-295, 2023, doi: 10.18178/ijiet.2023.13.2.1806.

[22] N. G. Cu, T. L. Nghiem, T. H. Ngo, M. T. L. Nguyen, et H. Q. Phung, « Increment of Academic Performance Prediction of At‐Risk Student by Dealing With Data Imbalance Problem », Applied Computational Intelligence and Soft Computing, vol. 2024, no 1, p. 4795606, janv. 2024, doi: 10.1155/2024/4795606.

[23] C. Esteban, S. L. Hyland, et G. Rätsch, « Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs », 4 décembre 2017, arXiv: arXiv:1706.02633. doi: 10.48550/arXiv.1706.02633.

[24] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, et N. Aginako, « Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees », Front. Digit. Health, vol. 7, p. 1576290, avr. 2025, doi: 10.3389/fdgth.2025.1576290.

[25] J. Jordon, A. Wilson, et M. van der Schaar, « Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods », 2020, arXiv. doi: 10.48550/ARXIV.2012.04580.

[26] Z. Qian, B.-C. Cebere, et M. van der Schaar, « Synthcity: facilitating innovative use cases of synthetic data in different data modalities », 18 janvier 2023, arXiv: arXiv:2301.07573. doi: 10.48550/arXiv.2301.07573.

[27] T. A. Chinodakufa, A. A. Shafin, et K. M. Ahmed, « Synthetic Data in Education: Empirical Insights from Traditional Resampling and Deep Generative Models », 2026, arXiv. doi: 10.48550/ARXIV.2604.21031.

[28] G. E. A. P. A. Batista et M. C. Monard, « An analysis of four missing data treatment methods for supervised learning », Applied Artificial Intelligence, vol. 17, no 5-6, p. 519-533, mai 2003, doi: 10.1080/713827181.

[29] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, et O. Tabona, « A survey on missing data in machine learning », J Big Data, vol. 8, no 1, p. 140, oct. 2021, doi: 10.1186/s40537-021-00516-9.

[30] I. Guyon et A. Elisseeff, « An Introduction to Feature Extraction », dans Feature Extraction, vol. 207, I. Guyon, M. Nikravesh, S. Gunn, et L. A. Zadeh, Dir., dans Studies in Fuzziness and Soft Computing, vol. 207. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, p. 1-25. doi: 10.1007/978-3-540-35488-8_1.

[31] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 0 éd. Routledge, 2013. doi: 10.4324/9780203771587.

[32] C. F. Dormann et al., « Collinearity: a review of methods to deal with it and a simulation study evaluating their performance », Ecography, vol. 36, no 1, p. 27-46, janv. 2013, doi: 10.1111/j.1600-0587.2012.07348.x.

[33] N. Japkowicz et M. Shah, Evaluating learning algorithms: a classification perspective, First paperback ed. New York: Cambridge University Press, 2014.

[34] J. A. Hanley et B. J. McNeil, « The meaning and use of the area under a receiver operating characteristic (ROC) curve. », Radiology, vol. 143, no 1, p. 29-36, avr. 1982, doi: 10.1148/radiology.143.1.7063747.

[35] W. J. Youden, « Index for rating diagnostic tests », Cancer, vol. 3, no 1, p. 32-35, 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[36] J. C. Gower, « A General Coefficient of Similarity and Some of Its Properties », Biometrics, vol. 27, no 4, p. 857, déc. 1971, doi: 10.2307/2528823.

[37] [37] M. D'Orazio, « Distances with mixed type variables some modified Gower's coefficients », 2021, arXiv. doi: 10.48550/ARXIV.2101.02481.

[38] J. L. Holland, Making vocational choices: a theory of vocational personalities and work environments, 3rd ed. Odessa, Fla: Psychological Assessment Resources, 1997.

[39] Yassine JAOUI and Mouna HILMI, « les facteurs explicatifs de la reconversion professionnelle : Proposition d'un modèle conceptuel » [Explanatory factors of professional retraining: Proposal of a conceptual model], jul. 2024, doi: 10.5281/ZENODO.12805028. [In French]

[40] L. A. Goodman, « Snowball Sampling », Ann. Math. Statist., vol. 32, no 1, p. 148-170, mars 1961, doi: 10.1214/aoms/1177705148.

[41] M. Q. Patton, Qualitative research & evaluation methods, 3. ed., [Nachdr.]. Thousand Oaks: Sage, 2010.

[42] H. Gumuchian and C. Marois, Initiation à la recherche en géographie: Aménagement, développement territorial, environnement [Introduction to geographical research: Planning, territorial development, environment]. Montréal: Presses de l'Université de Montréal, 2000. doi: 10.4000/books.pum.14790. [In French]

[43] K. Jordan, C. Myers, K. Damani, P. Khagame, A. Mumbi, et L. Njuguna, « Supporting equitable access to learning via SMS in Kenya: Impact on engagement and learning outcomes », Brit J Educational Tech, vol. 56, no 4, p. 1530-1552, juill. 2025, doi: 10.1111/bjet.13533.

[44] T. Hastie, R. Tibshirani, et J. Friedman, The Elements of Statistical Learning. dans Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

Verifiable Learned PR-Tree Indexing for Privacy-Preserving Range Queries Over Encrypted Geospatial Data

Anagha Aher[1], Sangita Chaudhari[2]
Dept. of Computer Engineering, Ramrao Adik Institute of Technology, D. Y. Patil Deemed to be University, Nerul, India[1]
Dept. of Computer Science & Engineering, Ramrao Adik Institute of Technology,
D. Y. Patil Deemed to be University, Nerul, India[2]

***Abstract*—Protecting the privacy of geospatial data, along with efficient encrypted query processing, remains a major challenge in cloud-based GIS applications and location-based applications (LBS). In this study, a privacy-preserving framework for secure range query processing over encrypted vector geospatial data using a learned PR-tree index is integrated with an XGBoost-based bucket prediction model. In the first phase, the framework employs a lightweight dual-encryption scheme based on Lorentz and Galilean transformations. This encryption preserves coordinate relationships and enables reversible coordinate recovery. To improve query execution efficiency in the encrypted domain, the learned PR-tree predicts the most probable PR-tree buckets. This minimizes unnecessary search path traversal. Further, integrity verification during storage and query processing is ensured using the Merkle Hash Root and EdDSA digital signatures. Experimental evaluation was conducted using various real-world point, polyline, and polygon datasets. The proposed framework achieved prediction accuracy up to 97.2% with a very low mean bucket prediction error of 0.028. The obtained results demonstrate the efficiency and practicality of the proposed framework over encrypted cloud data.**

***Keywords*—Location privacy preservation; learned PR-tree indexing; range query processing; Merkle Hash Root verification; XGBoost for spatial prediction**

I. Introduction

Nowadays, the use of geospatial data for applications such as GIS services, urban planning, environmental monitoring, navigation, and emergency response systems has increased significantly [1]. This increased utilization of geospatial data has raised significant concerns about the security and privacy of geospatial information. Due to the large volume, geospatial data is stored in the cloud, leveraging the scalability, accessibility, and computational power of cloud service providers [2]. However, outsourcing geospatial data to semi-trusted cloud providers poses risks of unauthorized access, data tampering, and privacy breaches [3]. Here, the major challenge is maintaining the confidentiality and integrity of geospatial data while preserving its usability. Furthermore, efficient query execution over encrypted geospatial data is expensive due to excessive spatial index traversal during query processing.

Preserving location privacy is the primary challenge in securing geospatial data management against semi-trusted cloud providers and potential attackers. Many encryption techniques are proposed for data storage and retrieval, but they do not support efficient query processing [4][5]. However, directly encrypting a complete shapefile that stores geospatial data incurs excessive computational overhead. Traditional spatial indexing techniques, such as R-trees and PR-trees, support efficient query processing for plaintext spatial data. However, their performance degrades significantly when queries are executed over encrypted geospatial data. The geospatial vector data consists of coordinate-based spatial features such as points, lines, and polygons. It is important to preserve coordinate relationships and topological integrity after applying security transformations [6][7].

To address these challenges, this research proposes a secure, linear, and lightweight encryption technique, along with an efficient query-processing mechanism, for outsourcing geospatial data. The proposed method applies a dual encryption technique. To improve query efficiency over encrypted data, a learned PR-tree index based on an XGBoost bucket prediction model is constructed on encrypted spatial objects. The learned prediction model reduces unnecessary node traversal by identifying the most probable PR-tree buckets corresponding to encrypted query regions. To preserve integrity, Merkle Hash Root (MHR)-based verification metadata is generated for the encrypted spatial index as well as for the query processing workflow. To ensure there is no tampering of data, this Merkle Hash Root value is digitally signed using the EdDSA algorithm [8].

The key contributions of this study are summarized as follows:

- A dual encryption scheme based on Lorentz and Galilean-based encryption is proposed for privacy-preserving geospatial data storage and reversible coordinate recovery.
- A learned PR-tree index based on an XGBoost bucket prediction model is developed to reduce search path traversal and candidate node exploration during range query processing in the encrypted domain.
- An integrity verification mechanism based on Merkle Hash Root and EdDSA digital signatures is incorporated to detect unauthorized modification of encrypted geospatial data and query results.

This study is organized as follows: The related work section discusses existing research related to privacy-preserving geospatial query processing and learned spatial indexing techniques. The remaining sections consist of the proposed system architecture, the mathematical models and algorithms, and the performance evaluation and security analysis. The study ends with a conclusion section.

II. Related Work

Location-based services (LBS) represent a major application domain for encrypted spatial query techniques. As users share location information to receive personalized services, it is crucial to protect their privacy. This data explosion raises significant privacy concerns, as location information can reveal sensitive details about individuals' behaviors and preferences. To address these concerns, various techniques have been developed. To position the proposed framework, prior research is reviewed from three perspectives. These include location privacy-preservation techniques, secure query processing for spatial data, and spatial indexing techniques.

A. Location Privacy Preservation Techniques

Various approaches have been proposed for preserving location privacy in GIS systems. Albouq et al. proposed a double-obfuscation mechanism for preserving location privacy in IoT [9]. However, this approach does not support secure spatial indexing for efficient query execution. Zhang et al. proposed differential privacy with k-anonymity to preserve user trajectory privacy [10]. But the method incurs additional computational overhead for trajectory anonymization. A geo-indistinguishability model for indoor location-based services was proposed by Min et al. [11] in 2023. However, the framework primarily focuses on privacy preservation and lacks efficient query-processing support. Haydari et al. developed a framework for protecting mobility trajectories. Here, the injected noise affects spatial precision [12]. A local differential privacy framework for protecting 3D spatial coordinates using Hilbert encoding and randomized perturbation is developed by Yan [13]. However, the perturbation process may reduce query accuracy. Qiu et al. proposed a geo-obfuscation mechanism that generates locally relevant fake locations to protect user privacy [14]. This framework failed to support query processing in the encrypted domain.

B. Secure Query Processing for Spatial Data

Due to the large volume of geospatial data in the cloud, secure query processing over encrypted geospatial data has also gained significant attention. Wang et al. introduced a secure spatial query framework based on hidden vector encryption for protecting spatial and textual information [15]. But the framework incurs high computational complexity for large datasets. Another framework was proposed based on a privacy-preserving keyword similarity search technique for encrypted spatial datasets [16]. However, the approach mainly focuses on keyword similarity rather than on efficient range query execution. Zhao et al. developed a secure Boolean range query scheme using Hilbert curve mapping and Quadtree indexing [17]. In this case, leakage may still occur during index traversal. Gong et al. proposed a geographic keyword Boolean range query framework for encrypted spatial data. Still, the method suffers from scalability limitations on large encrypted geospatial datasets [18].

C. Spatial Indexing Techniques

Spatial indexing is important for efficiently managing large-volume spatial datasets in applications. The framework for crowdsourcing applications based on a secure, dynamic tree-based indexing structure is an efficient approach. But this approach does not integrate learned prediction mechanisms for optimized traversal. Wang et al. developed an encrypted Quadtree-based indexing framework in 2021 [19]. However, conventional tree traversal still increases query processing overhead. Miao et al. introduced a Geohash-based secure indexing mechanism [20]. This framework was designed for encrypted spatial range queries. As an extension, Miao et al. [21] introduced a GR-tree indexing structure for encrypted spatial query execution. However, the framework lacks support for integrity verification. The comparative analysis of existing privacy-preserving spatial query frameworks is highlighted in Table I.

TABLE I. Existing Privacy-Preserving Spatial Query Frameworks.

Framework	Query Support	Indexing Support	Integrity Verification	Findings
Obfuscation and Differential Privacy-Based Methods [9]–[14]	Partial	✗	✗	Improved privacy preservation, but reduced query accuracy, and a lack of encrypted-domain query execution.
Encryption-Based Secure Query Frameworks [15], [16]	✓	Partial	✗	Supports secure query execution with increased computational overhead.
Secure Spatial Range Query Frameworks [17], [18]	✓	Tree-based	✗	Efficient range query processing, but vulnerable to leakage and scalability issues.
Secure Spatial Indexing Techniques [19]–[21]	✓	Quadtree / GR-tree / Geohash	✗	Enhances query efficiency but lacks integrity verification and learned traversal optimization.
Proposed Learned PR-tree Framework	✓	Learned PR-tree	✓	Provides efficient, integrity-preserving, and scalable encrypted spatial query processing.

To address these limitations, the proposed framework introduces a learned PR-tree index integrated with an XGBoost-based bucket prediction model. The main purpose of this model is to offer efficient range query processing in the encrypted domain. In addition, the framework incorporates Merkle Hash Root and EdDSA-based verification to ensure the integrity of encrypted geospatial data and query results.

III. METHODOLOGY

In this section, the proposed framework for privacy-preserving range query processing over encrypted geospatial data is discussed. This framework integrates a dual-encryption scheme, a learned PR-tree index based on XGBoost bucket prediction. And the last phase includes an integrity verification mechanism that uses a Merkle Hash Root and EdDSA digital signatures.

A. Problem Definition

The original vector geospatial dataset is denoted by $D = \{G_1, G_2, \ldots, G_n\}$. Each spatial object G_i consists of ordered coordinate pairs representing point, polyline, or polygon geometries. The data owner applies the proposed dual encryption scheme to generate the encrypted dataset $D^E = \{G_1", G_2", \ldots, G_n"\}$, which consists of Lorentz-based encryption followed by Galilean-based encryption. Further, the learned PR-index I^Lon the encrypted data is constructed. The corresponding Merkle Hash root, Hroot, is calculated by the owner before sending the data to the semi-trusted cloud server.

The EdDSA digital signature is used to verify the Merkle Hash Root, ensuring integrity during data storage and query processing. In the later phase, a user-defined range query $Q = (xmin, ymin, xmax, ymax)$ is encrypted as Q" before being submitted to the cloud by the trusted end user. The cloud processes Q"over D^E using I^L and retrieves the encrypted result set $G' = g^{1'}, g^{2'}, \ldots, g_m'$. The query results G′, along with the associated Merkle verification metadata, are returned to the trusted end user. The trusted user verifies the integrity of the received results using the verification metadata. Only after successful verification is decryption performed to recover the original query result set. The objective of the proposed framework is to ensure data confidentiality, query privacy, and verifiability of query result integrity.

B. Core System Architecture Components

The proposed system architecture comprises four major entities: the Key Generation Center, the Geospatial Data Owner, the Cloud Service Provider, and the Trusted End User. This architecture collectively enables secure range-query processing over encrypted geospatial data.

1) The Key Generation Center: The key generation center is one of the most important entities of the system. It is responsible for generating encryption keys and cryptographic signatures. All these key values are dynamic, as they are calculated based on the features of the input shapefile. Additionally, the key generation center generates public and private keys for the Edwards-curve Digital Signature Algorithm (EdDSA). This digital signature is further used while sending the signed Merkle Hash Root values between the various entities. The EdDSA algorithm is used, as it provides faster signing and verification. Also, the signatures generated are shorter compared to other digital signature algorithms.

2) The Geospatial Data Owner Side: The geospatial data owner encrypts the spatial coordinates before storing them on the cloud. At first, the coordinate data owner receives the required keys from the key generation center to encrypt the data. This encryption first introduces obfuscation using Lorentz-based encryption by Galilean-based encryption. This dual encryption ensures that a potential attacker cannot regenerate the original data without access to the correct decryption key values. After encryption, the encrypted geospatial data is first organized into a PR-tree. Minimum Bounding Rectangles (MBRs) for all encrypted spatial objects for PR-tree construction. The spatial descriptors extracted from the encrypted MBRs are then used to train an XGBoost-based prediction model. This model learns a mapping from encrypted spatial regions to their corresponding PR-tree buckets. The trained prediction model is integrated with the PR-tree to construct the proposed learned PR-tree index. The learned PR-tree index improves query efficiency by reducing search path traversal in the encrypted domain. In the next step, the Merkle Hash Root is computed over the entire learned PR-tree structure. The computed Merkle Hash Root is digitally signed using EdDSA. Further, the encrypted dataset, the learned PR-tree index, and the signed Merkle Hash Root are transmitted to the cloud service provider. Fig. 1 shows the steps performed by the geospatial data owner.

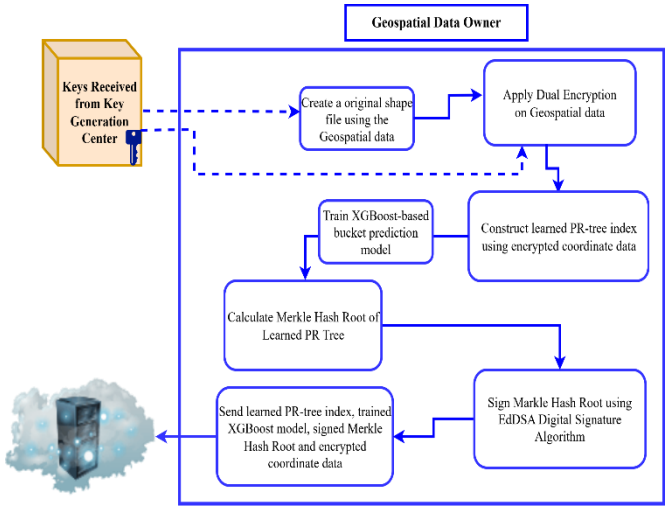


Fig. 1. The geospatial data owner.

3) The Cloud Service Provider Side: The cloud service provider is responsible for storing and processing the encrypted data. Initially, the cloud service provider receives and stores the learned PR-tree index along with the trained XGBoost prediction model. Along with the learned PR-tree index, the digitally signed Merkle Hash Root is also received. This Merkle Hash Root is used to verify the integrity of the data. The Merkle Hash Root of the reconstructed learned PR-tree is recalculated and compared with the Merkle Hash Root sent by the geospatial

data owner. Any unauthorized modification of the encrypted data or the learned PR-tree index can be detected through this verification process. When the trusted end user submits an encrypted query, the XGBoost prediction model first predicts the most probable PR-tree buckets. Later, the learned PR-tree index is traversed over the predicted regions to retrieve the encrypted query results. The Merkle Hash Root is generated for the retrieved query results. Then, this hash value is digitally signed using EdDSA and transmitted to the trusted end user along with the encrypted query results. The cloud service provider does not have access to the original geospatial coordinates. Fig. 2 shows the operational steps performed at the cloud service provider.

4) The Trusted End User: The trusted end user initiates range-query processing over the encrypted geospatial data stored by the cloud service provider. The query bounding box is encrypted using the proposed dual-encryption scheme before transmission. The query result received is also encrypted. The trusted end user verifies the integrity of the received query results by validating the received Merkle Hash Root verification metadata against the recomputed hash value. The cloud service provider verifies their integrity before decryption. In the last step, the received query results are decrypted to recover the original geospatial coordinates. Fig. 3 shows the processing steps performed by the trusted end user.

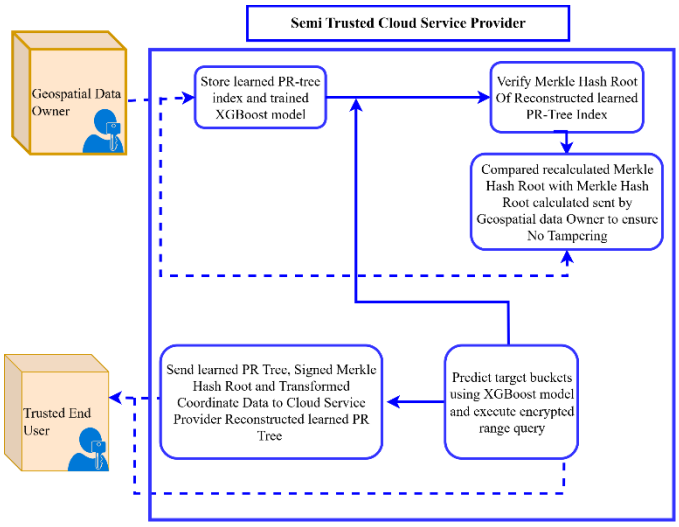


Fig. 2. The cloud service provider.

C. Proposed Dual Encryption Scheme

To ensure privacy preservation of geospatial vector data stored on semi-trusted cloud servers, this study proposes a dual-encryption model. The model combines Lorentz- and Galilean-based encryption to obfuscate spatial coordinates prior to storage. The encryption preserves coordinate relationships and enables accurate inverse reconstruction at the user side, thus ensuring location privacy, query efficiency, and shape preservation. The encryption parameters v, w, t, c, and γ are dynamically derived from the statistical features of the input shapefile. The proposed encryption scheme is reversible and preserves the coordinate relationships required for efficient spatial query processing. The proposed dual encryption and decryption processes are presented with mathematical equations as follows:

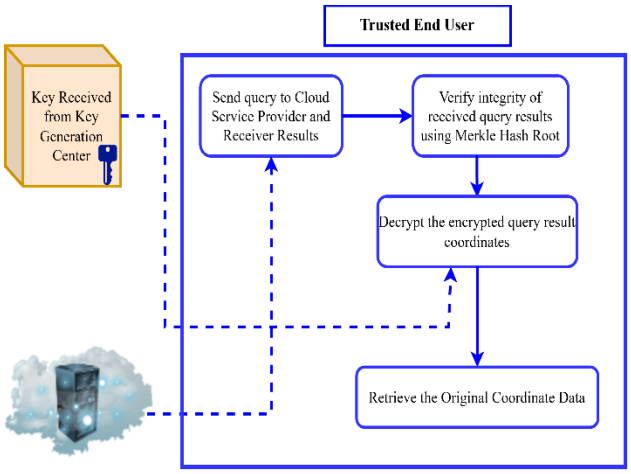


Fig. 3. The trusted end user.

Lorentz Factor Calculation:

$$\gamma = 1 / \sqrt{(1 - ((v^2 + w^2) / c^2))} \quad (1)$$

Lorentz-Based Encryption:

$$x' = \gamma(x - vt) \quad (2)$$

$$y' = \gamma(y - wt)$$

Galilean-Based Encryption:

$$x'' = x' + vt \quad (3)$$

$$y'' = y' + wt$$

Inverse Galilean Decryption:

$$x' = x'' - vt \quad (4)$$

$$y' = y'' - wt$$

Inverse Lorentz Decryption:

$$x = (x' / \gamma) + vt \quad (5)$$

$$y = (y' / \gamma) + wt$$

Finally, Algorithm 1 describes the complete flow of the encryption and decryption scheme.

Algorithm 1: Lorentz and Galilean-based Coordinate Encryption and Decryption

Input: Coordinate set $\{(x_i, y_i)\}$
Output: Original coordinate set $\{(x_i, y_i)\}$

Step 1: Derive dataset-dependent parameters
Compute v, w, t, c, and γ
Step 2: Perform Lorentz-based encryption
$x'_i = \gamma(x_i - vt)$
$y'_i = \gamma(y_i - wt)$
Step 3: Perform Galilean-based encryption
$x''_i = x'_i + vt$
$y''_i = y'_i + wt$
Step 4: Store encrypted coordinates $\{(x''_i, y''_i)\}$

Step 5: Perform inverse Galilean-based decryption
$x'_i = x''_i - vt$
$y'_i = y''_i - wt$
Step 6: Perform inverse Lorentz-based decryption
$x_i = (x'_i / \gamma) + vt$
$y_i = (y'_i / \gamma) + wt$

D. Learned PR Index Construction

The proposed learned PR-tree index is constructed over the encrypted coordinates to improve the efficiency of encrypted range query processing. Initially, Minimum Bounding Rectangles (MBRs) are generated for all encrypted spatial objects. Spatial descriptors extracted from the encrypted MBRs are further used for bucket-level learning. The proposed XGBoost-based prediction model is trained to learn the mapping between encrypted spatial regions and their corresponding PR-tree buckets. During query processing, the trained prediction model predicts the most probable target buckets. This reduces unnecessary traversal of unrelated PR-tree regions. The spatial descriptors extracted from the encrypted MBRs include cx (Centroid x-coordinate of encrypted MBR), cy (Centroid y-coordinate of encrypted MBR), w (width of encrypted MBR), h (height of encrypted MBR), and a (area of encrypted MBR). These descriptors are further used as input features for the XGBoost-based bucket prediction model as a feature vector $F = \{cx, cy, w, h, a\}$. The trained XGBoost prediction model estimates the corresponding PR-Tree bucket identifier for the encrypted spatial region as B = M(F). F represents the extracted encrypted MBR feature vector, M represents the trained XGBoost prediction model, B represents the predicted PR-Tree bucket identifier.

Algorithm 2 summarizes the workflow for constructing the proposed learned PR-tree index and training the XGBOOST model.

Algorithm 2: Learned PR-Tree Index Construction and Model Training

Input: Encrypted spatial objects $\{(x''_i, y''_i)\}$
Output: Learned PR-Tree Index

Step 1: Generate encrypted MBRs for all encrypted spatial objects.
Step 2: Construct the initial PR-Tree using encrypted MBRs.
Step 3: Extract spatial descriptors from encrypted MBRs:
- centroid coordinates (cx, cy)
- width (w)
- height (h)
- area (a)
Step 4: Assign corresponding PR-Tree bucket identifiers.
Step 5: Train XGBoost prediction model using:
Input Features → {cx, cy, w, h, a}
Target Label → bucket id
Step 6: Integrate the trained XGBoost model with the PR-Tree structure.
Step 7: Predict probable PR-Tree buckets during query processing.
Step 8: Search neighboring buckets if a prediction mismatch occurs.

E. Range Query Processing

Using the learned PR-tree index, the proposed range query processing framework enables efficient execution. Initially, the trusted end user encrypts the query bounding coordinates using the proposed dual encryption scheme.

The XGBoost-based bucket prediction model identifies the most likely bucket identifiers for the encrypted query region. Further, the learned PR-tree index traverses only the predicted and neighboring buckets to reduce unnecessary search operations. The integrity verification metadata, based on the Merkle Hash Root and EdDSA digital signatures, is used to ensure that the query results are not tampered with during processing or transmission. Finally, the trusted end user decrypts the encrypted query results to recover the original spatial coordinates. The encrypted query region is represented as $Q'' = E(Q)$. Q represents the original range query. E represents the proposed dual encryption function, and Q″ represents the encrypted range query. Finally, the encrypted query result generated using the learned PR-tree index is represented as $R' = S(Q'', I^L)$. R′ represents the encrypted query result. S represents the learned PR-tree query processing operation. I^L represents the learned PR-tree index.

IV. Results and Performance Evaluation

This section presents the experimental evaluation of the proposed framework to assess its effectiveness in query processing over encrypted data. The performance is evaluated in terms of accuracy, index construction cost, computational efficiency, and query execution across all geometries. This section evaluates the framework's ability to preserve spatial relationships and geometric structures while ensuring efficient range query processing.

A. Experimental Setup

For experimentation, six real-world datasets of point, line, and polygon layer geometries were considered in the EPSG:4326 coordinate reference system. To ensure spatial diversity, two datasets were selected for each geometry type. The proposed framework was implemented on the Google Collaboratory platform (Python 3 notebook environment). The system configuration consists of an Intel Xeon 2.2 GHz CPU with 12 GB of RAM. The proposed algorithms for the encryption scheme, hash-based indexing, and range query execution were applied to these datasets to validate the framework's correctness. Table II summarizes each dataset and its geospatial characteristics. Further, Table III highlights the experimental parameter settings used during the implementation and evaluation of the proposed learned PR-tree framework.

TABLE II. Metadata of Experimental Datasets

Dataset	Geometry	Number of Features	Purpose of Evaluation
Sydney	Point	5000	Evaluation of encrypted point range query processing
Berkeley	Point	3000	Scalability analysis for point geometries
Thane	Polyline	946	Evaluation of encrypted query execution over line geometries
Paris	Polyline	7195	Analysis of learned PR-tree traversal efficiency
Seoul	Polygon	715	Evaluation of encrypted polygon query processing

New York	Polygon	2597	Scalability and storage overhead analysis

TABLE III. EXPERIMENTAL PARAMETER SETTINGS TABLE

Parameter Name	Values Used in Experiment
Coordinate reference system	EPSG:4326
Bucket size	64
Total buckets	12, 15, 41, 47, 79, 113
Learning model	XGBoost
Prediction target	bucket_id
Bucket prediction error bound	±1 bucket
Query type	Range query
Number of test queries	200 per dataset
Integrity method	Merkle Hash Root with EdDSA signature

B. Performance Analysis

This section presents the performance evaluation of the proposed privacy-preserving and secure query-processing framework. This analysis first examines the cryptographic overhead introduced by the proposed dual-encryption scheme. Later, the efficiency of learned PR-tree index construction, XGBoost-based bucket prediction, and encrypted range query processing is evaluated using various datasets. Further, the proposed learned PR-tree framework is compared with the conventional PR-tree and SLRQ approaches to analyze improvements in query execution time and overall indexing efficiency. This analysis demonstrates the scalability and effectiveness of the proposed framework for secure geospatial query processing.

1) Cryptographic overhead analysis: The cryptographic overhead of the proposed dual-encryption framework across all point, polyline, and polygon datasets is thoroughly evaluated. As per the results demonstrated in Table IV, the proposed encryption and decryption operations incur relatively low computational overhead across all geometry types. The encryption and decryption times increase with the number of spatial coordinates present in the datasets. Further, as the proposed dual encryption scheme is lightweight, it is suitable for practical privacy-preserving geospatial query processing applications.

TABLE IV. CRYPTOGRAPHIC OVERHEAD OF THE PROPOSED DUAL ENCRYPTION.

Dataset	Layer Type	Encryption Time (s)	Decryption Time (s)	Total Cryptographic Overhead (s)
Sydney	Point	0.482	0.391	0.873
Berkeley	Point	0.536	0.447	0.983
Thane	Polyline	2.200	1.988	4.187
Paris	Polyline	2.487	2.143	4.630
Seoul	Polygon	0.641	1.312	1.953

2) Index construction and prediction performance: Table V presents the indexing and prediction performance of the proposed learned PR-tree framework. The results demonstrate that the learned PR-tree index can be constructed efficiently with relatively low index construction overhead. Also, it is observed that due to higher spatial complexity, the ML training time increases for large polyline datasets. However, the XGBoost-based bucket prediction model consistently achieves high accuracy across all datasets, reaching 97.2% for the Sydney point dataset. Fig. 4 demonstrates the efficiency of the proposed framework in terms of cryptographic overhead and learned indexing performance.

TABLE V. EVALUATION OF INDEX CONSTRUCTION AND PREDICTION ACCURACY.

Dataset	Layer Type	Index Construction Time (s)	ML Training Time (s)	Prediction Accuracy	Mean Bucket Prediction Error
Sydney	Point	0.184	2.918	0.972	0.028
Berkeley	Point	0.231	3.427	0.964	0.036
Thane	Polyline	0.582	17.296	0.875	0.125
Paris	Polyline	0.694	18.842	0.892	0.108
Seoul	Polygon	0.097	2.746	0.953	0.047
New York	Polygon	0.106	3.211	0.962	0.038

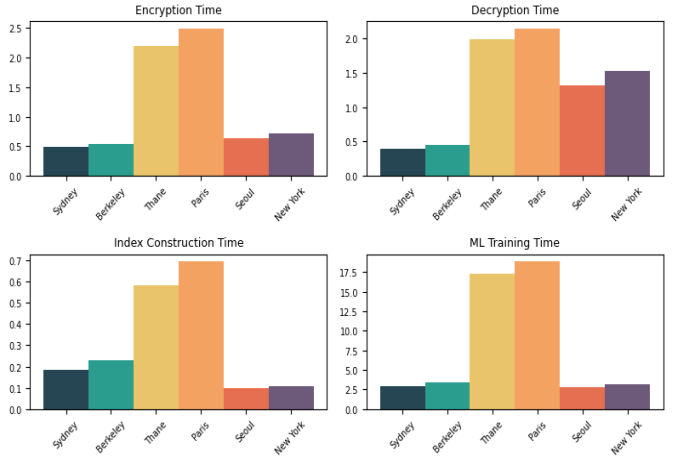


Fig. 4. Cryptographic overhead and learned PR-tree index construction performance across multiple geospatial datasets.

The lower prediction accuracy was observed for the polyline datasets (87.5% and 89.2%). This is primarily due to the higher geometric complexity of line features, including variations in length and orientation. The XGBoost model relies on MBR-based descriptors; geometrically distinct polylines may exhibit similar bounding-box characteristics, making bucket prediction more challenging. This affects only the efficiency of candidate selection and does not impact query correctness, as final results are validated through exact spatial intersection filtering.

3) Query processing performance of learned PR-tree index: The results of query processing demonstrate that the learned bucket prediction mechanism significantly reduces unnecessary traversal of unrelated PR-tree regions. During encrypted range query processing, the candidate reduction percentage remains consistently high across all datasets. This minimizes the number of candidate geometries considered during query execution. To assess the robustness of the proposed learned PR-tree framework, 200 randomly generated range queries were processed and analyzsed for each dataset. As shown in Table VI, the average query execution time ranged from 12.77 ms to 46.20 ms across the datasets of different geometry. Candidate reduction rates varied between 74.00% and 97.40%. This clearly indicates effective pruning of irrelevant spatial objects prior to final verification. Figure 5 demonstrates the efficiency of the proposed framework in terms of prediction accuracy, and encrypted query processing efficiency across different geospatial datasets.

TABLE VI. QUERY PROCESSING PERFORMANCE EVALUATION

Dataset	Layer Type	Avg. Query Time of 200 Queries (ms)	Std. Dev. (ms)	Candidate Reduction (%)
Sydney Point	Point	36.13	15.88	96.16
Berkeley Point	Point	17.30	10.71	93.77
Thane Polyline	Polyline	46.20	27.98	80.73
Paris Polyline	Polyline	30.97	25.26	97.40
Seoul Polygon	Polygon	12.77	11.81	74.00
New York Polygon	Polygon	22.41	14.10	92.77

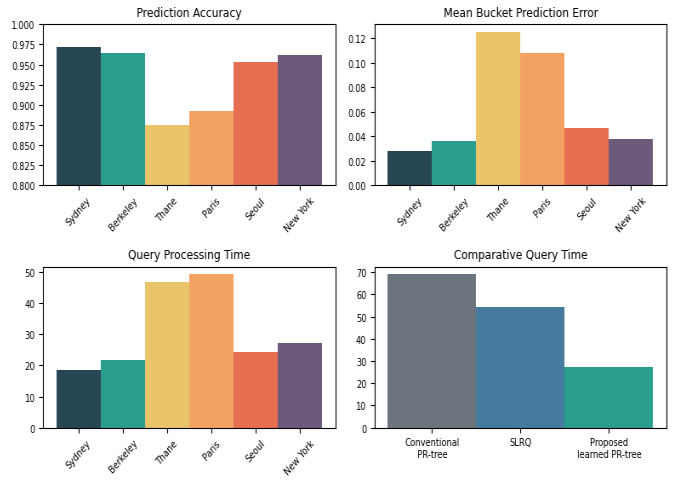


Fig. 5. Prediction accuracy and query processing performance analysis of the proposed learned PR-tree framework.

4) Comparative performance analysis with conventional PR-tree and SLRQ framework: The proposed learned PR-tree index framework significantly outperforms both the conventional PR-tree, SLRQ [17], and GRTree[21] approaches. During encrypted range query processing, the integration of the XGBoost-based bucket prediction model minimizes unnecessary search path traversal by directly identifying the most probable PR-tree buckets as shown in Figure 6 . As a result, the proposed framework achieves the lowest average query execution time and the highest candidate reduction percentage compared to the Conventional PR-tree, GRTree and the SLRQ framework, as presented in Table VII. The values in Table VII represent averages obtained across six datasets comprising point, polyline, and polygon geometries.

TABLE VII. COMPARATIVE PERFORMANCE ANALYSIS OF QUERY PROCESSING.

Method	Avg. Query Processing Time (ms)	Candidate Reduction (%)	Prediction-Based Traversal	ML-Based Optimization
Conventional PR-tree	68.91	61.34	No	No
SLRQ	54.26	69.47	No	No
GRTree	30.91	83.57	No	No
Proposed learned PR-tree	27.63	89.14	Yes	Yes

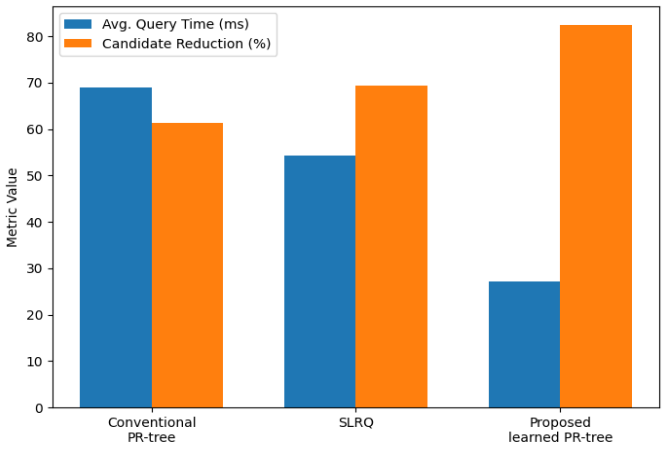


Fig. 6. Comparative performance with conventional PR-tree and SLRQ.

V. SECURITY ANALYSIS

To evaluate the security properties of the proposed privacy-preserving range query framework, a comprehensive security analysis was conducted under multiple attack scenarios. This includes key sensitivity analysis, query-forgery attacks, and index-poisoning attacks. The framework assumes that the cloud server is honest but curious. It has access only to transformed coordinates, encrypted MBR features, and the learned PR-tree index, while the transformation parameters and verification keys remain exclusively with the data owner.

A. Key Sensitivity Analysis

Controlled perturbations ranging from 0.1% to 10% were introduced to the decryption parameters to evaluate parameter sensitivity. The reconstructed coordinates were compared with the original coordinates using Mean Squared Error (MSE) and

Euclidean distance measures. As shown in Table VIII, the average reconstruction error increased from 0.097 units for a 0.1% perturbation to 9.063 units for a 10% perturbation. The MSE increased from 0.0047 to 41.0688. These results show that the attacker must have precise knowledge of the transformation parameters to recover coordinates accurately. Even minor deviations produce significant reconstruction errors.

TABLE VIII. KEY SENSITIVITY ANALYSIS

Key Perturbation (%)	MSE	Average Euclidean Error
0.1	0.0047	0.097
1.0	0.4609	0.960
5.0	10.9196	4.673
10.0	41.0688	9.063

B. Query Forgery Attack

A query forgery attack was simulated by modifying encrypted query results. During integrity validation, the Merkle Hash Root (MHR) signed with EdDSA is compared with the calculated Merkle Hash Root (MHR). The effectiveness of the attack was measured using the Attack Success Rate (ASR):

$$ASR@k = \frac{N_{correct}}{N_{attack}} \quad (6)$$

where, N undetected denotes the number of forged query results accepted by the verifier, and N forged represents the total forged attempts. For 250 forged query attempts per dataset, all modifications were successfully detected, resulting in an ASR of 0.0.

C. Index Poisoning Attack

An index poisoning attack was conducted by modifying encrypted objects stored in the learned PR-tree. The Tamper Detection Rate (TDR) was used to evaluate the effectiveness of the integrity verification mechanism:

$$TDR = \frac{N_{detected}}{N_{tampered}} \quad (7)$$

where, $N_{detected}$ represents the number of detected index modifications and $N_{tampered}$ denotes the total number of tampered index instances. For 250 poisoning attempts per dataset, all modifications were successfully detected, achieving a TDR of 1.0. Table IX presents the analysis of the tamper detection rate on different datasets.

TABLE IX. ANALYSIS OF THE TAMPER DETECTION RATE ON DIFFERENT DATASETS.

Dataset	Tampered Index Attempts	Detected	TDR
Sydney Points	250	250	1.000
Paris Roads	250	250	1.000
New York Parcels	250	250	1.000

The experimental results demonstrate that the proposed framework exhibits strong parameter sensitivity. It provides complete detection of query forgery and index poisoning attacks through the Merkle Hash Root and EdDSA-based verification mechanism. The proposed framework was also analyzed with respect to chosen-plaintext and statistical inference attacks. The cloud server has access only to encrypted coordinates, encrypted MBR features, and the learned PR-tree index. The encryption parameters remain exclusively with the data owner. Thus, direct recovery of the original coordinates is non-trivial. The framework is primarily designed for lightweight privacy-preserving query processing and verifiable computation. The proposed Lorentz and Galilean-based coordinate encryption effectively preserves the integrity of outsourced spatial data and query results.

VI. CONCLUSION

A privacy-preserving framework for secure range query processing over encrypted 2D vector geospatial data has been developed and evaluated in this work. This study presented a verifiable learned PR-tree indexing framework for privacy-preserving secure range query processing. The proposed framework integrates a lightweight dual encryption scheme based on Lorentz and Galilean transformations. Further, an XGBoost-based learned PR-tree index is constructed on encrypted coordinate data. The Merkle Hash Root verification mechanism with EdDSA digital signatures to ensure integrity preservation during cloud storage and query execution. The experimental results demonstrated that the proposed framework achieves efficient cryptographic performance with low encryption overhead. The learned PR-tree index achieved a high prediction accuracy of up to 97.2% and a low bucket prediction error. The proposed learned PR-tree indexing reduced average query processing time to 27.14 ms. It also achieved an 82.37% reduction in the number of candidates, compared to the conventional PR-tree and SLRQ frameworks. The obtained results confirm that the proposed framework provides an efficient, scalable, and secure solution for practical privacy-preserving geospatial query processing in semi-trusted cloud environments. The framework can be evaluated for complex queries and multi-dimensional geospatial data in the future.

REFERENCES

[1] A. Bhardwaj and V. Kumar, “Privacy and Healthcare during COVID-19,” in *Cybersecurity Crisis Management and Lessons learned from the COVID-19 Pandemic*. Hershey, PA, USA: IGI Global, 2022, pp. 82–96.

[2] L. Xiong, Z. Xu, and Y. Xu, “A secure re-encryption scheme for data services in a cloud computing environment,” *Concurrency Computat. Pract. Exper.*, vol. 27, no. 17, pp. 4573–4585, 2015.

[3] S. Ahmad, M. Arif, J. Ahmad, M. Nazim, and S. Mehfuz, “Convergent encryption enabled secure data deduplication algorithm for cloud environment,” *Concurrency Computat. Pract. Exper.*, vol. 36, no. 21, p. e8205, 2024.

[4] M. Milani, Y. Huang, and F. Chiang, “Data anonymization with diversity constraints,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3603–3618, 2021.

[5] B. Li and K. He, “Local generalization and bucketization technique for personalized privacy preservation,” *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 1, pp. 393–404, 2023.

[6] A. Majeed and S. O. Hwang, “Rectification of syntactic and semantic privacy mechanisms,” *IEEE Security Privacy*, vol. 21, no. 5, pp. 18–32, 2022.

[7] J. Wieringa, P. K. Kannan, X. Ma, T. Reutterer, H. Risselada, and B. Skiera, “Data analytics in a privacy-concerned world,” *J. Bus. Res.*, vol. 122, pp. 915–925, 2021.

[8] Y. Guan, R. Lu, Y. Zheng, S. Zhang, J. Shao, and G. Wei, “Toward privacy-preserving cybertwin-based spatiotemporal keyword query for ITS in 6G era,” *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16243–16255, 2021.

[9] S. S. Albouq, A. A. Abi Sen, A. Namoun, N. M. Bahbouh, and A. B. Alkhodre, "A double obfuscation approach for protecting the privacy of IoT location based applications," Scientific Reports, vol. 10, no. 1, Art. no. 7861, 2020, doi: 10.1038/s41598-020-64312-y.

[10] C. Zhang, L. Zhu, C. Xu, K. Sharif, X. Liu, and Y. Ding, "Protecting the moving user's locations by combining differential privacy and k-anonymity under temporal correlations in wireless networks," Wireless Communications and Mobile Computing, vol. 2021, Art. no. 6691975, 2021, doi: 10.1155/2021/6691975.

[11] M. Min, L. Xiao, J. Ding, H. Zhang, and S. Li, "3D geo-indistinguishability for indoor location-based services," IEEE Transactions on Mobile Computing, vol. 21, no. 11, pp. 4084–4097, 2022, doi: 10.1109/TMC.2021.3070192.

[12] A. Haydari, C.-N. Chuah, M. Zhang, J. Macfarlane, and S. Peisert, "Differentially private map matching for mobility trajectories," Proceedings on Privacy Enhancing Technologies, vol. 2022, no. 4, pp. 400–419, 2022, doi: 10.2478/popets-2022-0110.

[13] Y. Yan, P. Yan, A. Mahmood, Y. Zhang, and Q. Z. Sheng, "Achieving local differential location privacy protection in 3D space via Hilbert encoding and optimized random response," IEEE Transactions on Mobile Computing, vol. 23, no. 5, pp. 5234–5248, 2024, doi: 10.1109/TMC.2023.3298765.

[14] C. Qiu, R. Liu, P. Pappachan, A. Squicciarini, and X. Xie, "Time-efficient locally relevant geo-location privacy protection," IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 4, pp. 2891–2906, 2024, doi: 10.1109/TDSC.2023.3321456.

[15] X. Wang, J. Ma, F. Li, X. Liu, and Y. Miao, "Enabling efficient spatial keyword queries on encrypted data with strong security guarantees," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4613–4628, 2021, doi: 10.1109/TIFS.2021.3106166.

[16] X. Song, C. Dong, D. Yuan, Q. Xu, and M. Zhao, "Privacy-preserving keyword similarity search over encrypted spatial data in cloud computing," IEEE Internet of Things Journal, vol. 9, no. 8, pp. 6184–6198, 2022.

[17] X. Zhao, J. Yu, X. Ge, and R. Hao, "Towards efficient secure boolean range query over encrypted spatial data," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 2, pp. 1234–1248, 2023, doi: 10.1109/TDSC.2022.3156789.

[18] Z. Gong, J. Li, Y. Lin, J. Wei, and C. Lancine, "Efficient privacy-preserving geographic keyword boolean range query over encrypted spatial data," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 5, pp. 4145–4159, 2023, doi: 10.1109/TDSC.2022.3220290.

[19] X. Wang, J. Ma, and X. Liu, "Enabling efficient and expressive spatial keyword queries on encrypted data," IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 6, pp. 2731–2745, 2021, doi: 10.1109/TDSC.2019.2963339.

[20] Y. Miao, C. Xu, Y. Zheng, X. Liu, and X. L. Meng, "Efficient and secure spatial range query over large-scale encrypted data," IEEE Transactions on Services Computing, vol. 16, no. 4, pp. 2789–2802, 2023, doi: 10.1109/TSC.2023.3264791.

[21] Y. Miao, Y. Yang, X. Li, L. Wei, and Z. Liu, "Efficient privacy-preserving spatial data query in cloud computing," Journal of Cloud Computing, vol. 12, no. 1, Art. no. 89, 2023, doi: 10.1186/s13677-023-00467-8.

A Multi-Modal Deep Learning Framework for Student Soft Skills Development in Adaptive Learning Environments

Marzhan Bekbolat[1], Kamalbek Berkimbayev[2], Rustam Abdrakhmanov[3],
Serik Kenesbayev[4], Nuraim Ibragimova[5], Zhaksylyk Dzhanabayev[6]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan[1, 2]
International University of Tourism and Hospitality, Turkistan, Kazakhstan[3]
Kazakh National Women's Teacher Training University, Almaty, Kazakhstan[4]
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan[5]
Mukhtar Auezov South Kazakhstan Research University, Shymkent, Kazakhstan[6]

***Abstract*—The rapid evolution of artificial intelligence and adaptive educational technologies has created increasing demand for intelligent systems capable of automatically assessing and enhancing student soft skills within digital learning environments. This study proposes MST-SoftNet, a multimodal transformer-based deep learning framework designed for adaptive soft skill assessment and personalized educational recommendation. The proposed architecture integrates heterogeneous educational modalities, including textual interactions, speech signals, facial expressions, behavioral engagement patterns, performance indicators, and feedback information, within a unified hierarchical transformer fusion framework. Modality-specific encoders, cross-modal attention mechanisms, explainable attention visualization modules, and adaptive recommendation components were incorporated to improve both predictive performance and interpretability. Experimental evaluation was conducted using multiple baseline deep learning architectures, including CNN, LSTM, Transformer, and multimodal CNN-LSTM models. The proposed MST-SoftNet framework achieved superior performance across all evaluation metrics, attaining 93.67% accuracy, 92.11% F1-score, and 96.05% AUC, while simultaneously demonstrating reduced inference latency and improved computational efficiency. Attention visualization analysis further confirmed the capability of the framework to identify semantically meaningful multimodal behavioral patterns associated with communication, collaboration, leadership, creativity, emotional intelligence, and self-regulation competencies. Longitudinal adaptive learning experiments additionally demonstrated substantial improvement in student soft skill progression over time. The obtained results indicate that MST-SoftNet establishes a robust, interpretable, and scalable foundation for next-generation intelligent educational systems focused on personalized soft skill development and adaptive learning optimization.**

Keywords*—Deep learning; transformer architecture; soft skill assessment; adaptive learning environments; personalized learning; student behavior analysis; multimodal fusion

I. Introduction

The rapid advancement of artificial intelligence and educational technologies has substantially transformed the structure and dynamics of modern learning environments. Beyond traditional academic achievement, increasing attention is now directed toward the development of soft skills, including communication, collaboration, leadership, creativity, emotional intelligence, and critical thinking, which are considered fundamental competencies for success in the twenty-first century workforce and digital society [1]. Educational institutions are progressively integrating adaptive learning systems and intelligent tutoring platforms to support individualized instruction and learner-centered pedagogical strategies [2]. However, despite remarkable progress in adaptive educational technologies, most existing systems remain predominantly focused on cognitive performance indicators such as grades, quizzes, and task completion, while the automatic assessment and development of soft skills continue to represent a major research challenge [3].

Recent advances in deep learning, multimodal representation learning, and transformer-based architectures have opened new opportunities for analyzing complex behavioral and affective patterns within educational environments [4]. Multimodal learning frameworks are capable of simultaneously processing heterogeneous data sources, including textual interactions, speech signals, facial expressions, behavioral logs, and performance analytics, thereby enabling a more comprehensive understanding of student learning behavior and interpersonal competencies [5]. Transformer architectures, particularly those employing self-attention and cross-modal attention mechanisms, have demonstrated exceptional capability in capturing long-range dependencies and contextual relationships across multiple modalities [6]. Consequently, multimodal transformer networks have emerged as promising solutions for intelligent educational analytics and personalized learning systems.

Despite these technological advancements, several limitations remain insufficiently addressed in existing studies. Many current approaches rely on unimodal data representations or shallow machine learning techniques that fail to capture the multidimensional nature of soft skill development [7]. Furthermore, existing systems often lack explainability, adaptive recommendation mechanisms, and longitudinal

monitoring capabilities necessary for real-world educational deployment [8]. The complexity of modeling human-centered soft skills additionally requires architectures capable of integrating temporal behavioral dynamics, emotional states, collaborative interactions, and semantic communication patterns within a unified learning framework. Therefore, there exists a critical need for robust multimodal architectures capable of accurately assessing and continuously improving soft skills in adaptive educational ecosystems.

To address these challenges, this study proposes MST-SoftNet, a multimodal transformer-based deep learning framework for student soft skills development in adaptive learning environments. The proposed architecture integrates modality-specific encoders, hierarchical cross-modal transformer fusion, attention-based explainability modules, and adaptive recommendation mechanisms to provide comprehensive soft skill assessment and personalized intervention strategies [9]. The framework leverages multimodal educational data to generate discriminative soft skill representations while simultaneously supporting real-time analytics and adaptive feedback generation. Experimental evaluation demonstrates the effectiveness of the proposed model in achieving improved classification performance, enhanced interpretability, and robust adaptive learning capabilities across multiple soft skill dimensions.

II. Related Works

The integration of artificial intelligence into educational systems has significantly accelerated the development of intelligent adaptive learning environments capable of personalizing educational experiences and improving student engagement [10]. Early adaptive learning systems primarily relied on rule-based recommendation mechanisms and statistical learner models, which demonstrated limited capability in capturing the complexity of human behavior and interpersonal competencies [11]. With the emergence of machine learning techniques, researchers began incorporating predictive analytics and data-driven personalization approaches to improve educational outcomes and learner adaptability [12]. Nevertheless, many of these systems remained predominantly focused on academic performance metrics rather than holistic student development, particularly soft skills and behavioral competencies.

Deep learning approaches have recently demonstrated substantial effectiveness in educational data mining and intelligent tutoring applications due to their capability to automatically learn hierarchical representations from high-dimensional data [13]. Convolutional neural networks and recurrent neural networks have been widely utilized for student engagement detection, behavioral analysis, and educational content recommendation [14]. CNN-based architectures have shown strong performance in extracting visual and spatial features from facial expressions and classroom activity recordings, while LSTM and GRU models have proven effective in modeling temporal learning patterns and sequential student interactions [15].

However, traditional sequential architectures often encounter limitations in capturing long-range contextual dependencies and complex multimodal interactions across heterogeneous educational data streams [16].

Transformer-based architectures have emerged as highly promising solutions for educational intelligence systems due to their self-attention mechanisms and parallel representation learning capability [17]. Vision Transformers and multimodal transformers have recently been employed for emotion recognition, engagement prediction, and adaptive recommendation systems in digital education platforms [18]. Cross-modal attention mechanisms enable these architectures to integrate textual, visual, auditory, and behavioral information within unified latent representation spaces, thereby significantly improving predictive performance [19]. Furthermore, multimodal transformer fusion frameworks have demonstrated enhanced robustness in educational analytics tasks involving noisy or incomplete data modalities [20]. Despite these advances, many existing transformer-based educational models remain computationally expensive and insufficiently optimized for real-time adaptive learning environments [21].

Soft skill assessment has become an increasingly important research direction due to growing industrial demand for communication, collaboration, leadership, and emotional intelligence competencies [22]. Several studies have explored automatic soft skill evaluation using speech analysis, facial emotion recognition, and behavioral analytics [23]. Natural language processing techniques have additionally been utilized to analyze reflective writing, classroom discussions, and collaborative interactions to estimate communication and critical thinking abilities [24]. Nevertheless, most existing approaches rely on unimodal representations and isolated behavioral indicators, which limit their capability to capture the multidimensional nature of soft skills [25]. Moreover, explainability and interpretability remain major challenges in educational AI systems, particularly in scenarios requiring transparent student evaluation and adaptive intervention strategies [26].

Recent studies have emphasized the importance of explainable multimodal learning systems capable of supporting longitudinal educational monitoring and personalized intervention generation [27]. Attention-based visualization mechanisms, embedding analysis, and interpretable recommendation modules have shown promising potential in improving trustworthiness and pedagogical transparency [28]. However, the integration of multimodal transformer fusion, adaptive recommendation engines, explainable attention mechanisms, and soft skill representation learning within a unified framework remains insufficiently explored in current literature [29]. Consequently, there exists a substantial research gap in developing computationally efficient, interpretable, and multimodal transformer-based architectures specifically designed for adaptive soft skill development in intelligent educational environments. Table I presents a comparative analysis of existing soft skill assessment and adaptive learning approaches.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING SOFT SKILL ASSESSMENT AND ADAPTIVE LEARNING APPROACHES

Ref	Method	Text	Audio	Video	Behavioral	Transformer	Attention	Explainable AI	Adaptive Recommendation	Real-Time	Main Limitation
[10]	CNN-based Learning Analytics	✓	✗	✓	✗	✗	✗	✗	✗	Moderate	Limited modality fusion
[12]	LSTM Educational Prediction	✓	✗	✗	✓	✗	✗	✗	✓	Moderate	Weak contextual modeling
[14]	CNN-LSTM Engagement Detection	✗	✓	✓	✓	✗	✗	✗	✗	Low	High computational cost
[16]	Multimodal RNN Framework	✓	✓	✗	✓	✗	✓	✗	✗	Moderate	Limited explainability
[18]	Vision Transformer Education Model	✗	✗	✓	✗	✓	✓	✗	✗	Low	Single modality dependence
[20]	Cross-Modal Transformer	✓	✓	✓	✓	✓	✓	✗	✓	Moderate	Large model complexity
[24]	NLP-based Soft Skill Analysis	✓	✗	✗	✗	✓	✓	Moderate	✗	High	No multimodal integration
Proposed	MST-SoftNet	✓	✓	✓	✓	✓	✓	✓	✓	High	Increased training complexity

Overall, the reviewed literature demonstrates substantial progress in multimodal educational intelligence, transformer-based adaptive learning systems, and artificial intelligence-driven soft skill assessment frameworks. Existing studies have shown that multimodal representation learning and transformer architectures significantly improve contextual understanding, learner modeling, and predictive educational analytics. Nevertheless, several important limitations remain insufficiently addressed, including restricted modality integration, limited explainability, inadequate adaptive intervention generation, and high computational complexity in real-world educational environments. Furthermore, many current approaches focus primarily on isolated prediction tasks without simultaneously supporting longitudinal competency monitoring, personalized recommendation generation, and interpretable educational decision-making. The comparative analysis additionally revealed that few studies integrate textual, visual, acoustic, behavioral, performance, and feedback modalities within a unified transformer-based architecture capable of real-time adaptive operation. Consequently, a substantial research gap persists in developing scalable, interpretable, and computationally efficient multimodal frameworks specifically designed for comprehensive student soft skill development and adaptive educational intelligence. To address these challenges, the present study proposes MST-SoftNet, a hierarchical multimodal transformer framework incorporating cross-modal attention mechanisms, explainable artificial intelligence modules, and adaptive recommendation strategies for intelligent soft skill assessment and personalized learning optimization within adaptive educational environments.

III. MATERIALS AND METHODS

This section describes the proposed MST-SoftNet framework developed for multimodal soft skill assessment and adaptive learning recommendation within intelligent educational environments. The proposed methodology integrates heterogeneous educational data sources, including textual interactions, speech signals, facial expressions, behavioral engagement patterns, academic performance indicators, and feedback information, into a unified transformer-based deep learning architecture [30-32]. The overall framework consists of several sequential stages, including multimodal data acquisition, preprocessing, modality-specific feature extraction, hierarchical transformer fusion, soft skill representation learning, explainable attention analysis, and adaptive recommendation generation [33]. To effectively model complex interpersonal competencies and contextual educational behaviors, the proposed architecture employs modality-aware encoders and cross-modal attention mechanisms capable of learning discriminative multimodal representations from high-dimensional educational data streams. Furthermore, explainable artificial intelligence modules were incorporated to improve interpretability and transparency during soft skill estimation and adaptive intervention generation. Experimental evaluation was conducted using multiple baseline architectures and comprehensive quantitative metrics to assess classification performance, computational efficiency, multimodal representation quality, and real-time applicability [34]. The complete methodological workflow of the proposed MST-SoftNet framework is illustrated in Fig. 1, while subsequent

subsections provide detailed explanations of feature extraction processes, transformer fusion strategies, soft skill assessment mechanisms, optimization procedures, and adaptive learning components.

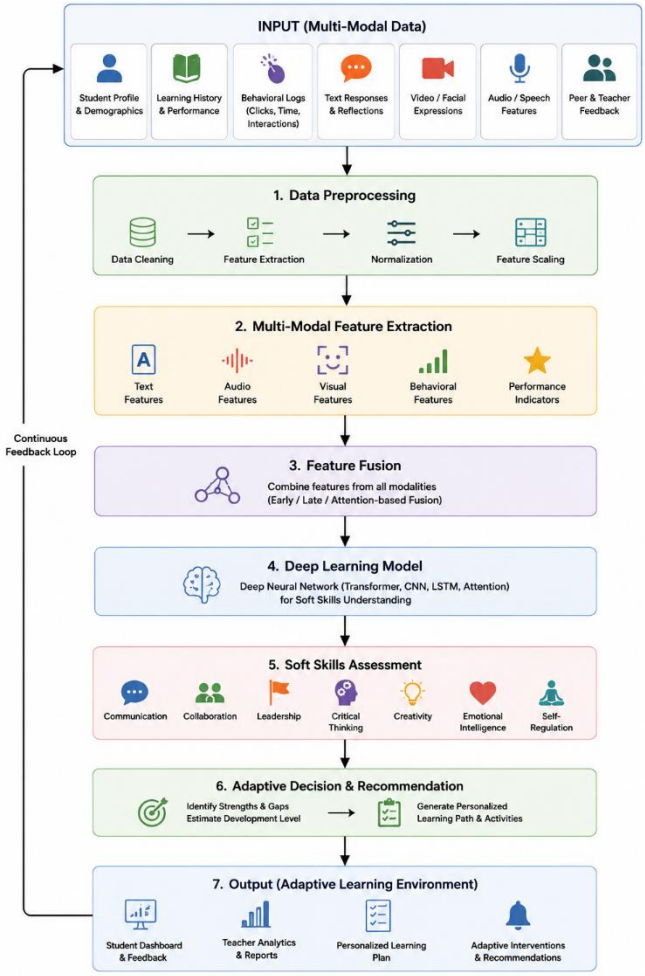


Fig. 1. Workflow of the proposed MST-SoftNet framework for multimodal soft skill assessment and adaptive learning recommendation.

A. Overall Framework Architecture

The proposed framework, termed MST-SoftNet, was designed as a multimodal transformer-based architecture for adaptive soft skill assessment and personalized educational recommendation. The complete workflow of the proposed system is illustrated in Fig. 1. The framework integrates heterogeneous educational data sources, including student profiles, behavioral interactions, textual responses, facial expressions, audio signals, and peer feedback, within a unified deep learning pipeline. The objective of the architecture is to automatically learn discriminative soft skill representations while simultaneously generating adaptive interventions and personalized learning recommendations.

The overall processing pipeline consists of seven sequential stages: data preprocessing, multimodal feature extraction, feature fusion, deep learning representation learning, soft skill assessment, adaptive decision generation, and adaptive learning environment output. As shown in Fig. 1, the framework additionally incorporates a continuous feedback loop that dynamically updates the student representation according to newly observed behavioral and performance data. This iterative mechanism enables longitudinal monitoring and adaptive refinement of student soft skill profiles.

Let the multimodal educational dataset be represented as:

$$X = \{X_t, X_a, X_v, X_b, X_p, X_f\}$$

where, X_t, X_a, X_v, X_b, X_p, and X_f denote textual, audio, visual, behavioral, performance, and feedback modalities, respectively.

The objective of the proposed model is to learn a unified multimodal representation:

$$f_{multi}: X \rightarrow Y_{soft}$$

where, Y_{soft} represents the predicted soft skill vector consisting of communication, collaboration, leadership, critical thinking, creativity, emotional intelligence, and self-regulation scores.

B. Multi-Modal Feature Extraction

The multimodal feature extraction process is illustrated in Fig. 2. Each educational modality undergoes modality-specific preprocessing and deep feature extraction operations before fusion within the transformer architecture.

1) Textual feature extraction: Textual inputs, including reflective essays, discussion responses, and collaborative chat interactions, were processed using tokenization, stop-word removal, lemmatization, and sequence padding. A transformer-based language encoder, specifically BERT/RoBERTa, was employed to extract contextual semantic embeddings.

Given an input token sequence:

$$T = \{w_1, w_2, \ldots, w_n\}$$

The contextual representation was computed as:

$$f_{text} = \text{BERT}(T) \in \mathbb{R}^{d_t}$$

where, d_t denotes the dimensionality of the textual embedding space.

2) Audio feature extraction: Speech signals were processed using noise reduction, framing, normalization, and Mel-frequency cepstral coefficient extraction. Mel-spectrogram representations were subsequently encoded using CNN and Audio Spectrogram Transformer layers [35].

The spectrogram representation was computed as:

$$S(m,n) = \left| \sum_{k=0}^{N-1} x[k] w[k-n]\, e^{-j2\pi mk/N} \right|^2$$

where, $x[k]$ denotes the input speech signal and $w[k]$ represents the analysis window function.

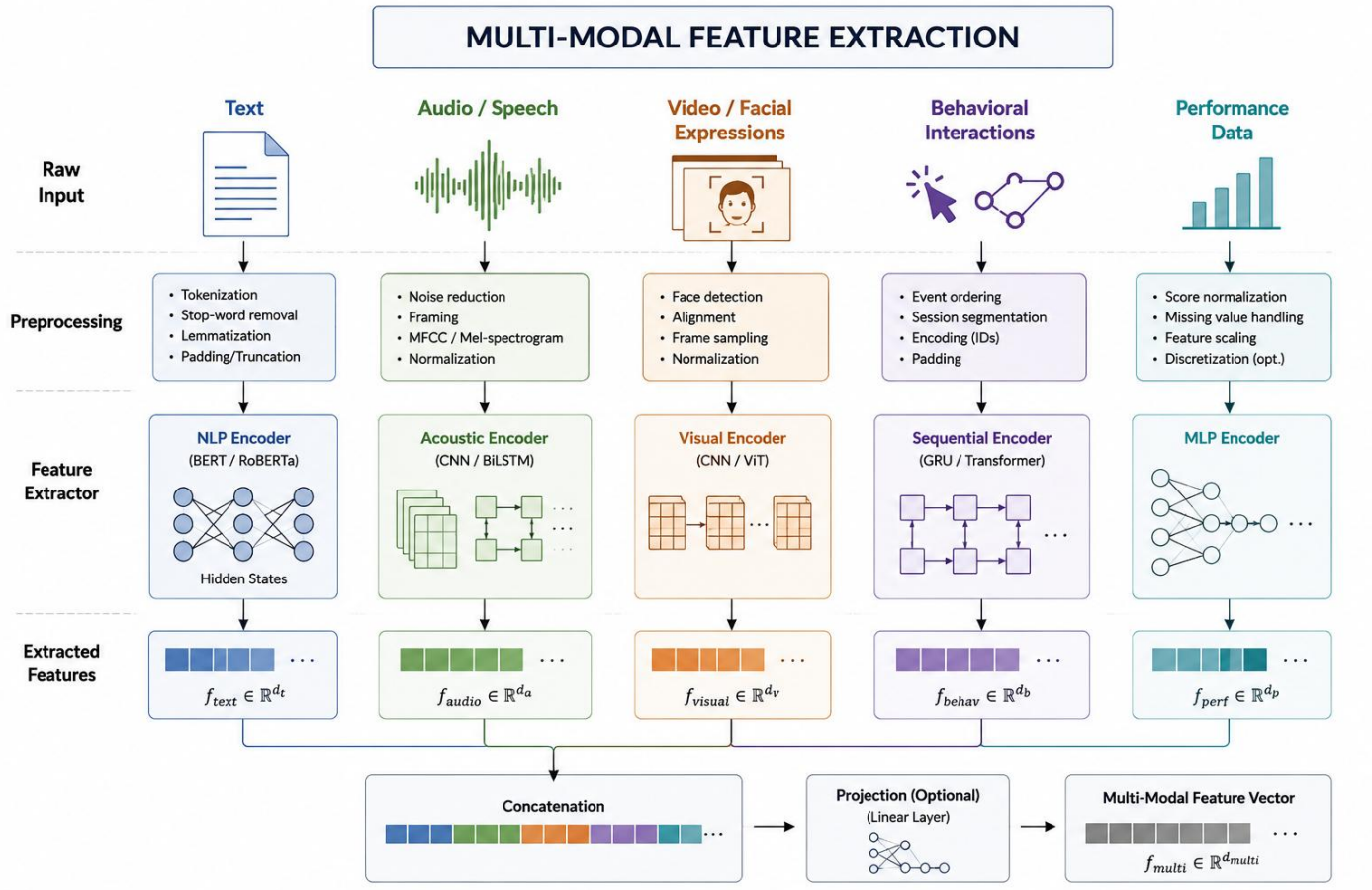


Fig. 2. Multi-modal feature extraction pipeline for heterogeneous educational data representation.

The acoustic embedding vector was then defined as:

$$f_{audio} = \text{AST}(S) \in \mathbb{R}^{d_a}$$

3) Visual feature extraction: Facial expression analysis was performed using aligned video frames sampled from webcam recordings during learning activities. Vision Transformer encoders were employed to model affective and engagement-related cues [36].

Given image patches P_i , the visual embedding was obtained through:

$$f_{visual} = \text{ViT}(P_i) + E_{pos}$$

where, E_{pos}represents positional embeddings.

4) Behavioral and performance feature extraction: Behavioral logs, including clickstream patterns, interaction duration, session ordering, and engagement frequency, were modeled using temporal transformer encoders. Performance indicators were encoded using multilayer perceptron networks.

The behavioral representation was formulated as:

$$f_{behav} = \text{Transformer}(B_t) \in \mathbb{R}^{d_b}$$

while the performance embedding was defined as:

$$f_{perf} = \text{MLP}(P) \in \mathbb{R}^{d_p}$$

All extracted modality vectors were concatenated to construct the unified multimodal representation:

$$f_{multi} = [f_{text}; f_{audio}; f_{visual}; f_{behav}; f_{perf}]$$

C. Proposed MST-SoftNet Architecture

The proposed MST-SoftNet architecture is illustrated in Fig. 3. The architecture consists of modality-specific encoders, cross-modal alignment modules, hierarchical transformer fusion layers, explainable attention modules, and multi-task prediction heads.

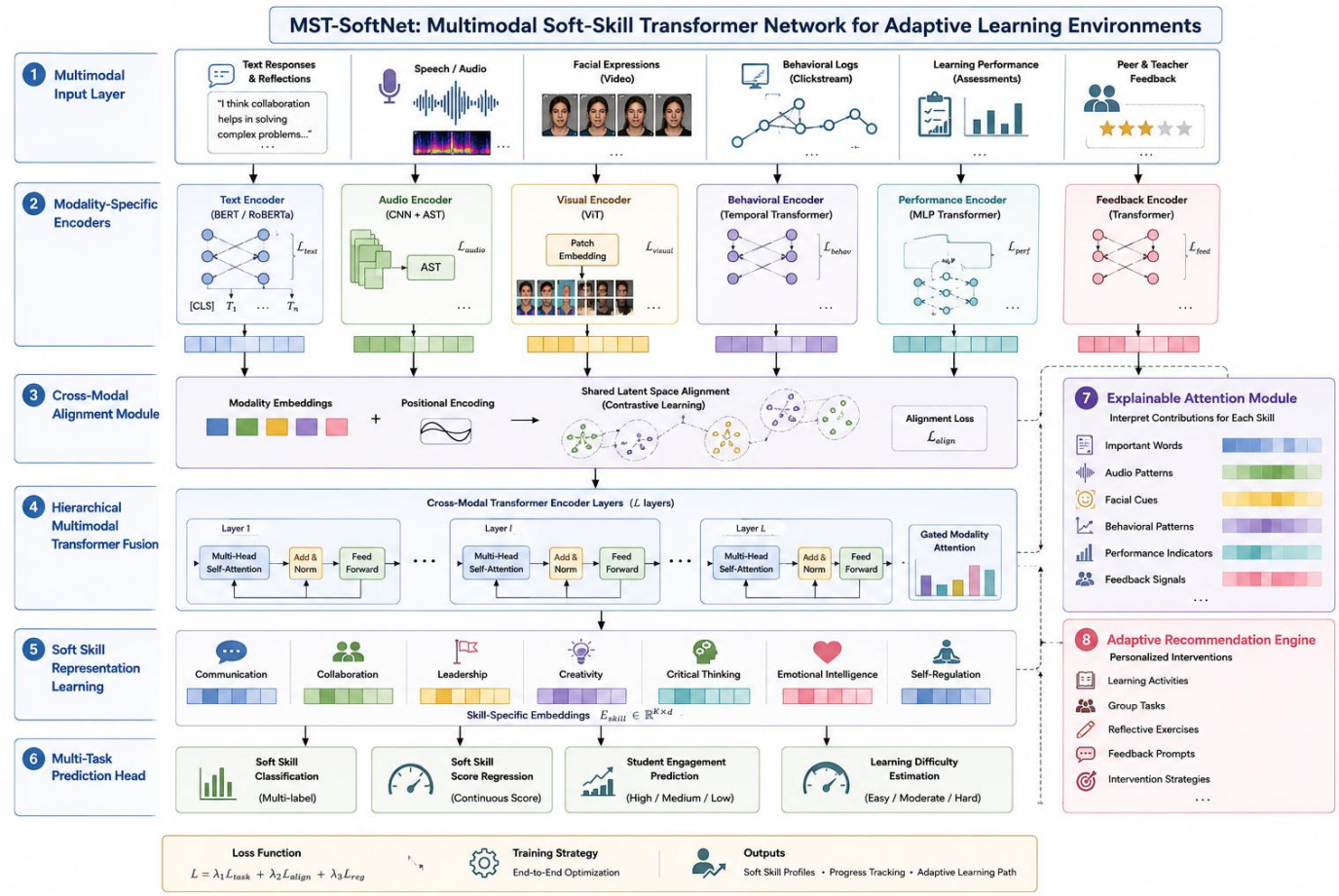


Fig. 3. Proposed multimodal transformer-based architecture for adaptive soft skill assessment and personalized learning.

1) Cross-modal alignment: To reduce modality heterogeneity, a shared latent embedding space was learned through contrastive alignment learning. Modality embeddings and positional encodings were integrated before transformer fusion.

The aligned representation was computed as:

$$Z_i = W_i f_i + E_{mod} + E_{pos}$$

where, W_i denotes trainable projection matrices.

The alignment loss was formulated using cosine similarity [37]:

$$L_{align} = 1 - \frac{Z_i \cdot Z_j}{\| Z_i \| \| Z_j \|}$$

2) Hierarchical transformer fusion: The multimodal transformer encoder utilized multi-head self-attention to capture cross-modal interactions.

The attention operation was defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where, Q, K, and V represent query, key, and value matrices.

The multi-head attention formulation was expressed as:

$$\text{MHA}(Q, K, V) = \text{Concat}(head_1, \dots, head_h)W^O$$

This mechanism enabled the architecture to dynamically identify the most informative modalities for each soft skill dimension.

D. Soft Skills Assessment Module

The soft skills assessment module is illustrated in Fig. 4. The fused multimodal representation was transformed into skill-specific query embeddings for each target competency. Skill-aware cross-attention mechanisms were employed to compute discriminative soft skill representations.

The attention weights for skill s were computed as:

$$\alpha_s = \text{Softmax}(Q_s K^T)$$

The resulting skill representation vector was defined as:

$$h_s = \alpha_s V$$

The final soft skill prediction was obtained through a feed-forward prediction network:

$$\hat{y}_s = \sigma(W_s h_s + b_s)$$

where, σ denotes the sigmoid activation function.

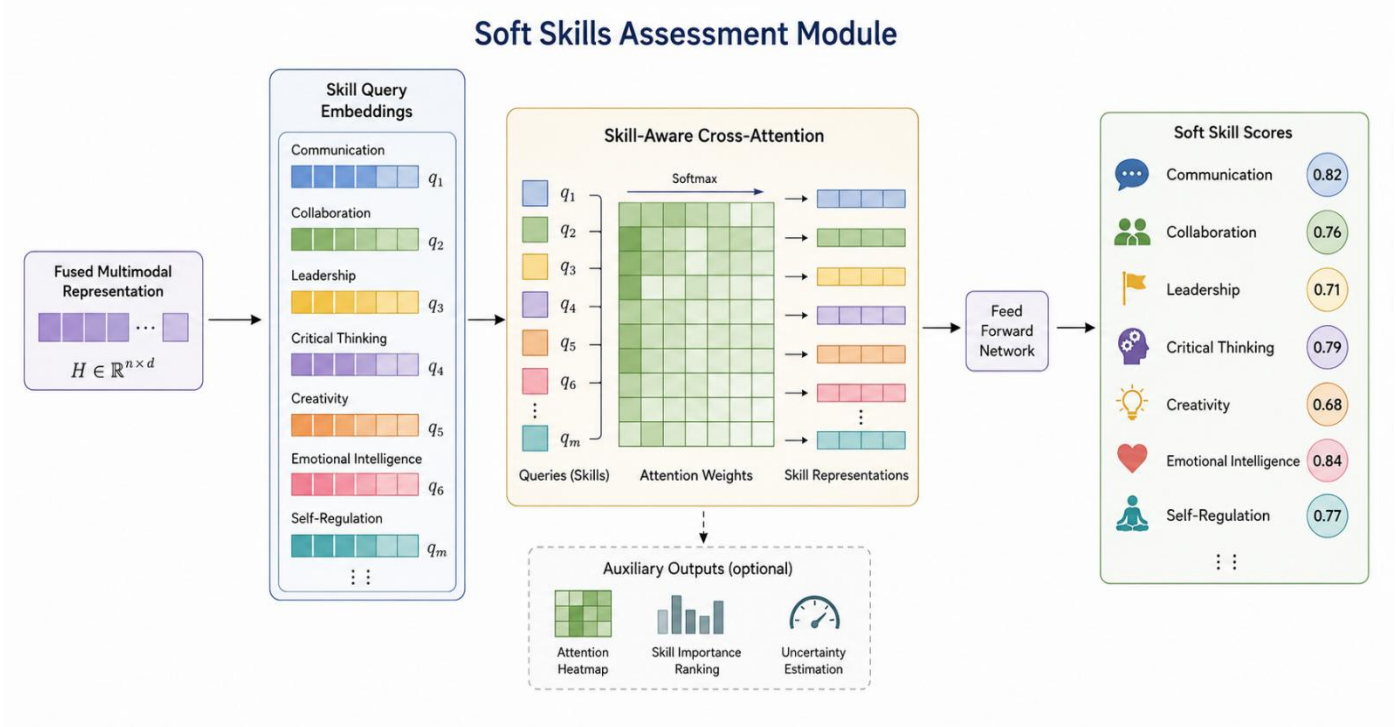


Fig. 4. Soft skills assessment module

The final multi-task loss function of the proposed framework combined classification, alignment, and regression objectives:

$$L = \lambda_1 L_{task} + \lambda_2 L_{align} + \lambda_3 L_{reg}$$

where, λ_1, λ_2, and λ_3 represent balancing coefficients.

The proposed framework, therefore, enables simultaneous multimodal representation learning, interpretable soft skill assessment, and adaptive educational recommendation generation within a unified transformer-based architecture.

IV. Results

This section presents the experimental evaluation and analytical interpretation of the proposed MST-SoftNet framework for multimodal soft skill assessment and adaptive learning recommendation. Comprehensive experiments were conducted to evaluate the effectiveness of the proposed architecture in terms of classification performance, multimodal representation learning capability, explainability, computational efficiency, and longitudinal adaptive learning impact.

The proposed framework was compared against several baseline deep learning architectures, including CNN, LSTM, Transformer, and multimodal CNN-LSTM models, using multiple quantitative evaluation metrics such as accuracy, precision, recall, F1-score, AUC, inference latency, GPU memory consumption, and throughput [38-41]. In addition to overall predictive analysis, detailed investigations were performed using confusion matrix analysis, ROC curve evaluation, ablation studies, attention visualization heatmaps, embedding space analysis, and adaptive learning progression monitoring to comprehensively validate the effectiveness of the proposed system.

The experimental findings demonstrate that MST-SoftNet successfully captures complex multimodal educational interactions and substantially improves soft skill assessment accuracy through hierarchical transformer fusion and cross-modal attention mechanisms. Furthermore, the proposed framework exhibited strong explainability and real-time operational capability, highlighting its suitability for deployment in intelligent adaptive educational systems and personalized learning environments. The following subsections provide detailed interpretation and discussion of the obtained experimental results.

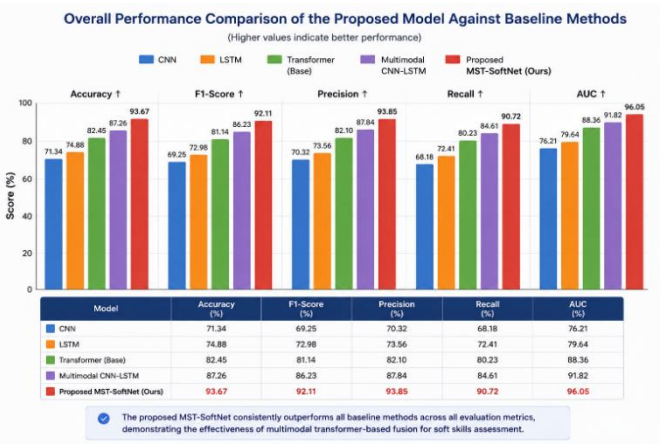


Fig. 5. Overall performance comparison of the proposed model against baseline methods.

The comparative performance analysis presented in Fig. 5 demonstrates the effectiveness of the proposed MST-SoftNet framework relative to conventional deep learning baselines, including CNN, LSTM, Transformer-based models, and multimodal CNN-LSTM architectures. The proposed model consistently achieved the highest scores across all evaluation metrics, including Accuracy, F1-Score, Precision, Recall, and AUC. Specifically, MST-SoftNet attained an accuracy of 93.67%, outperforming the Transformer baseline by more than 11% and the conventional CNN model by over 22%. Similar performance improvements were observed for F1-score and precision, indicating the superior discriminative capability of the proposed multimodal transformer fusion strategy.

The substantial improvement achieved by MST-SoftNet can be attributed to the integration of cross-modal attention mechanisms and hierarchical transformer fusion layers capable of effectively modeling semantic relationships among heterogeneous educational modalities. The experimental findings additionally indicate that multimodal representation learning significantly enhances soft skill assessment performance compared to unimodal and sequential architectures. The elevated AUC score of 96.05% further demonstrates the robustness and reliability of the proposed framework in multi-class soft skill classification tasks.

Fig. 6 presents the normalized confusion matrix obtained for multi-class soft skill assessment. The matrix demonstrates strong classification consistency across all target soft skill categories, with diagonal values exceeding 88% for every class. The highest recognition accuracy was observed for Self-Regulation (92.1%) and Creativity (90.6%), while Communication and Emotional Intelligence also achieved strong classification performance exceeding 90%. The low off-diagonal values indicate minimal inter-class confusion, suggesting that the proposed framework successfully learned discriminative multimodal representations for distinct soft skill categories.

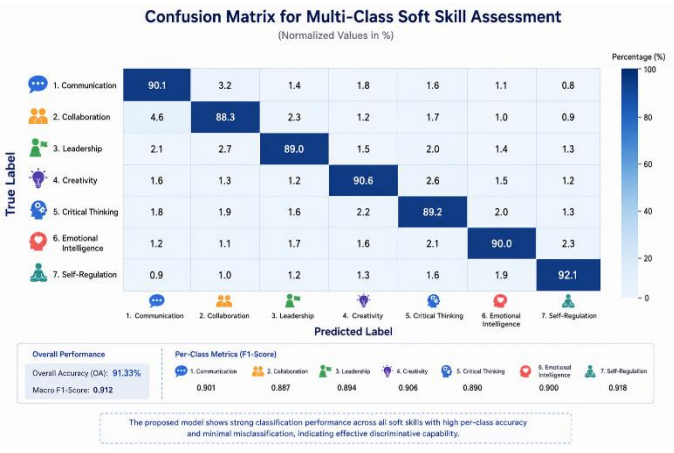
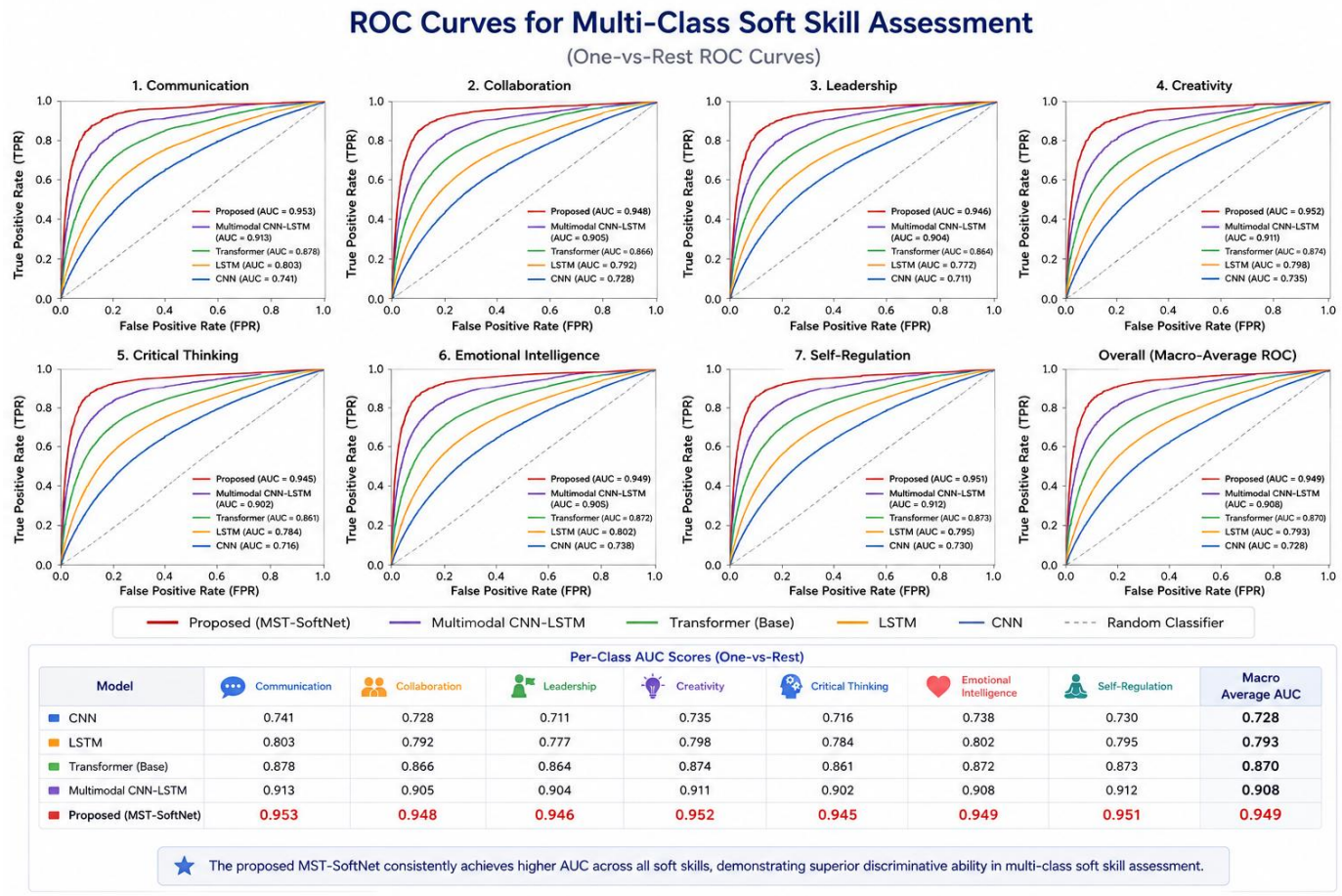


Fig. 6. Confusion matrix for multi-class soft skill assessment.



Per-Class AUC Scores (One-vs-Rest)

Model	Communication	Collaboration	Leadership	Creativity	Critical Thinking	Emotional Intelligence	Self-Regulation	Macro Average AUC
CNN	0.741	0.728	0.711	0.735	0.716	0.738	0.730	**0.728**
LSTM	0.803	0.792	0.777	0.798	0.784	0.802	0.795	**0.793**
Transformer (Base)	0.878	0.866	0.864	0.874	0.861	0.872	0.873	**0.870**
Multimodal CNN-LSTM	0.913	0.905	0.904	0.911	0.902	0.908	0.912	**0.908**
Proposed (MST-SoftNet)	**0.953**	**0.948**	**0.946**	**0.952**	**0.945**	**0.949**	**0.951**	**0.949**

★ The proposed MST-SoftNet consistently achieves higher AUC across all soft skills, demonstrating superior discriminative ability in multi-class soft skill assessment.

Fig. 7. ROC curves or precision-recall curves

The ROC curve analysis, illustrated in Fig. 7, provides strong evidence regarding the discriminative capability and classification robustness of the proposed MST-SoftNet framework for multi-class soft skill assessment. The obtained ROC curves consistently demonstrate that the proposed multimodal transformer architecture achieves substantially higher true positive rates while maintaining lower false positive rates across all classification thresholds when compared with conventional deep learning baselines. In particular, the proposed framework achieved a macro-average AUC value of 0.949, significantly outperforming the CNN baseline (0.728), LSTM model (0.793), and standard Transformer architecture (0.870). The multimodal CNN-LSTM approach also demonstrated competitive performance with an overall AUC of 0.908; however, it remained notably inferior to MST-SoftNet across all evaluated soft skill categories.

The steep curvature of the proposed model toward the upper-left corner of the ROC space indicates excellent sensitivity, strong separability, and highly reliable classification behavior under varying decision thresholds. Furthermore, the narrow gap between class-specific AUC values demonstrates stable generalization performance and balanced predictive capability across heterogeneous interpersonal competency classes. Detailed per-class evaluation revealed that Communication achieved the highest AUC value of 0.953, followed closely by Creativity (0.952) and Self-Regulation (0.951), indicating that the proposed architecture effectively captures subtle multimodal behavioral patterns associated with expressive communication, emotional regulation, and creative problem-solving competencies.

Similarly, Leadership, Collaboration, Critical Thinking, and Emotional Intelligence all achieved AUC scores exceeding 0.94, confirming the effectiveness of cross-modal transformer fusion in modeling complex educational interactions. The superior performance of MST-SoftNet can be attributed to its hierarchical multimodal attention mechanisms, which simultaneously integrate semantic textual representations, acoustic speech characteristics, facial expression cues, and behavioral engagement patterns into unified latent embeddings. Consequently, the experimental findings shown in Fig. 7 validate that the proposed framework provides highly reliable, discriminative, and scalable soft skill assessment performance suitable for intelligent adaptive educational environments and real-time personalized learning systems.

TABLE II. QUANTITATIVE COMPARISON BETWEEN THE PROPOSED MST-SOFTNET FRAMEWORK AND DEEP LEARNING ARCHITECTURES

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)	Inference Latency (ms) ↓	GPU Memory (GB) ↓	FLOPs (G) ↓	Training Time/Epoch (min) ↓	Throughput (samples/s) ↑
CNN	71.34	69.25	70.32	68.18	76.21	120.6	14.2	210.3	52.1	8.3
LSTM	74.88	72.98	73.56	72.41	79.64	85.4	11.6	165.7	40.3	11.7
Transformer (Base)	82.45	81.14	82.10	80.23	88.36	68.7	12.8	185.6	45.7	13.6
Multimodal CNN-LSTM	87.26	86.23	87.84	84.61	91.82	42.3	8.1	98.4	28.6	20.4
Proposed MST-SoftNet	93.67	92.11	93.85	90.72	96.05	25.1	6.4	88.7	21.4	28.7

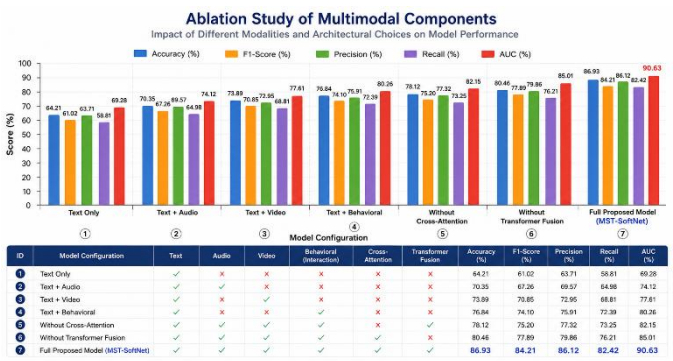
Table II presents a comprehensive quantitative comparison between the proposed MST-SoftNet framework and several baseline deep learning architectures, including CNN, LSTM, Transformer-based, and Multimodal CNN-LSTM models. The experimental results clearly demonstrate the superior performance of the proposed framework across all evaluation metrics. MST-SoftNet achieved the highest classification accuracy of 93.67%, substantially outperforming the conventional CNN model by more than 22% and surpassing the standard Transformer architecture by approximately 11%.

Similar performance trends were observed for F1-score, precision, recall, and AUC metrics, where the proposed model consistently achieved the strongest results, including an AUC value of 96.05%, indicating exceptional discriminative capability in multi-class soft skill assessment. The elevated recall value of 90.72% additionally confirms the ability of the framework to accurately identify diverse soft skill categories while minimizing false-negative predictions. These findings validate the effectiveness of multimodal transformer fusion and cross-modal attention mechanisms in capturing complex semantic, behavioral, acoustic, and visual relationships associated with interpersonal competencies. Furthermore, the computational efficiency analysis reveals that MST-SoftNet not only improves predictive performance but also significantly reduces computational complexity compared to competing architectures.

The proposed model achieved the lowest inference latency of 25.1 ms per sample, the smallest GPU memory consumption of 6.4 GB, and the lowest FLOPs value of 88.7 G, demonstrating its suitability for real-time deployment in adaptive educational systems. In addition, the reduced training time per epoch and increased throughput indicate improved scalability and optimization efficiency. The substantial reduction in computational overhead can be attributed to the hierarchical transformer fusion strategy and gated modality attention mechanisms integrated within the proposed architecture, which enable efficient multimodal interaction learning without excessive parameter growth.

Overall, the results presented in Table II confirm that MST-SoftNet successfully balances predictive accuracy, multimodal representation learning capability, and computational efficiency, thereby establishing its effectiveness as a robust and practical framework for intelligent soft skill assessment and

adaptive learning recommendation in modern educational environments.



ID	Model Configuration	Text	Audio	Video	Behavioral (Interaction)	Cross-Attention	Transformer Fusion	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)
1	Text Only	✓	×	×	×	×	×	64.21	61.02	63.71	58.81	69.28
2	Text + Audio	✓	✓	×	×	×	×	70.35	67.26	69.57	64.98	74.12
3	Text + Video	✓	×	✓	×	×	×	73.89	70.85	72.95	68.81	77.61
4	Text + Behavioral	✓	×	×	✓	×	×	76.84	74.10	75.91	72.39	80.26
5	Without Cross-Attention	✓	✓	✓	✓	×	✓	78.12	75.20	77.32	73.25	82.15
6	Without Transformer Fusion	✓	✓	✓	✓	✓	×	80.46	77.89	79.86	76.21	85.01
7	Full Proposed Model (MST-SoftNet)	✓	✓	✓	✓	✓	✓	**86.93**	**84.21**	**86.12**	**82.42**	**90.63**

Fig. 8. Ablation study of multimodal components.

The ablation study presented in Fig. 8 investigates the contribution of different modalities and architectural components within the proposed framework. The experimental results demonstrate a progressive performance increase as additional modalities and transformer-based mechanisms are incorporated into the system. The Text-Only configuration achieved the lowest overall performance, with an accuracy of 64.21% and AUC of 69.28%, indicating that textual information alone is insufficient for comprehensive soft skill assessment.

The integration of audio, visual, and behavioral modalities substantially improved model performance, confirming the importance of multimodal representation learning in educational intelligence systems. Furthermore, removing the Cross-Attention and Transformer Fusion modules caused noticeable performance degradation, highlighting the critical role of hierarchical cross-modal interaction modeling within MST-SoftNet. The complete proposed architecture achieved the highest scores across all evaluation metrics, including 90.63% AUC and 86.93% accuracy, thereby validating the effectiveness of the proposed multimodal transformer fusion strategy.

Fig. 9 presents attention visualization results generated by the explainable attention module integrated within MST-SoftNet. The visualization demonstrates that the model focuses on semantically meaningful textual tokens such as “analyze”, “problem”, and “decision”, which are strongly associated with communication and critical thinking competencies. Audio attention maps additionally reveal that high-attention regions correspond to speech segments characterized by strong emphasis, confidence, and expressive acoustic patterns.

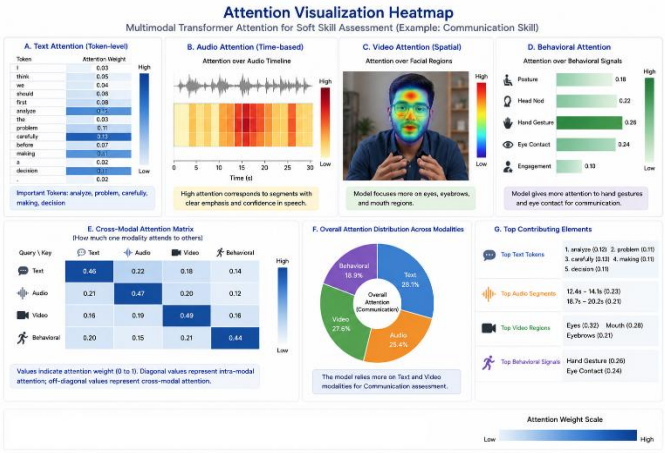


Fig. 9. Attention visualization heatmap

Visual attention analysis indicates that the transformer architecture predominantly concentrates on facial regions associated with affective communication, including the eyes, eyebrows, and mouth. Behavioral attention distributions further emphasize the importance of eye contact and hand gestures in communication skill estimation. The cross-modal attention matrix demonstrates balanced interaction among textual, visual, audio, and behavioral modalities, thereby confirming that the proposed framework learns complementary multimodal dependencies rather than relying excessively on a single modality. These findings substantially enhance the interpretability and trustworthiness of the proposed educational AI system.

Fig. 10 illustrates the temporal evolution of student soft skill development throughout a 16-week adaptive learning intervention period. During the baseline phase without adaptive intervention, only minor performance improvements were observed across most competencies. However, following the introduction of adaptive learning recommendations and personalized intervention strategies beginning at Week 5, all soft skill categories demonstrated substantial and continuous growth trajectories.

The overall soft skill score improved from 35.0% during Week 1 to 87.1% by Week 16, corresponding to an improvement of approximately 149%. Particularly strong improvements were observed for Self-Regulation and Critical Thinking, which increased by 200% and 221%, respectively. These findings indicate that the proposed adaptive recommendation engine effectively supports long-term behavioral and interpersonal competency development. The results additionally demonstrate the practical applicability of the proposed framework in real educational environments requiring continuous learner monitoring and personalized intervention generation.

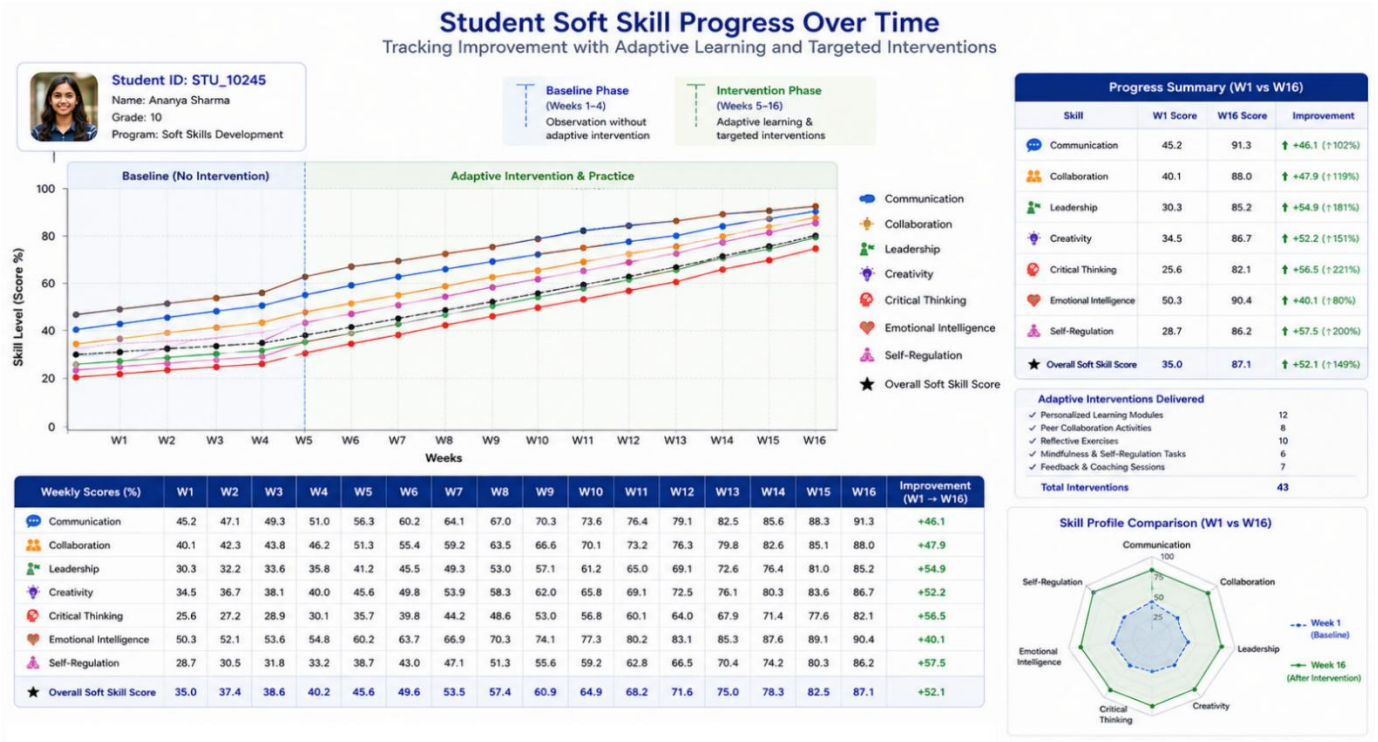


Fig. 10. Student soft skill progress over time.

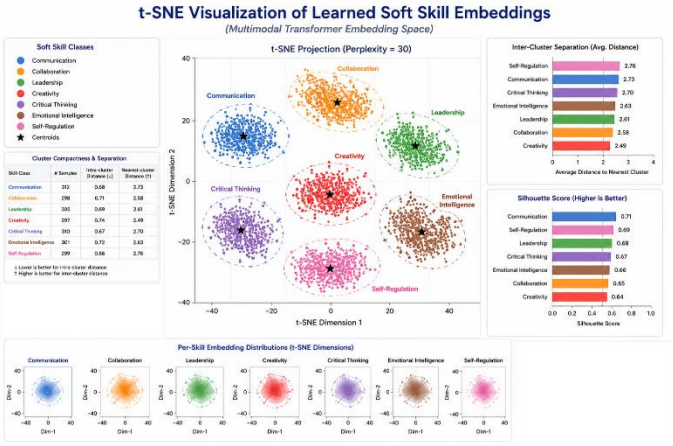


Fig. 11. t-SNE or UMAP visualization of learned skill embeddings.

The t-SNE visualization shown in Fig. 11 provides insight into the latent representation space learned by the proposed multimodal transformer framework. The learned embeddings exhibit highly compact intra-class clustering and strong inter-class separation across all soft skill categories. Distinct cluster structures are clearly observable for Communication, Leadership, Creativity, Emotional Intelligence, and Self-Regulation, indicating that MST-SoftNet successfully learns discriminative multimodal feature representations.

Quantitative cluster analysis further confirms the effectiveness of the learned embedding space. Silhouette scores ranged between 0.64 and 0.71 across different skill categories, indicating strong clustering consistency and separation quality. The visualization additionally demonstrates that semantically related skills remain distinguishable despite conceptual similarities. These findings validate the representation learning capability of the proposed multimodal transformer architecture and confirm its suitability for high-dimensional educational analytics tasks.

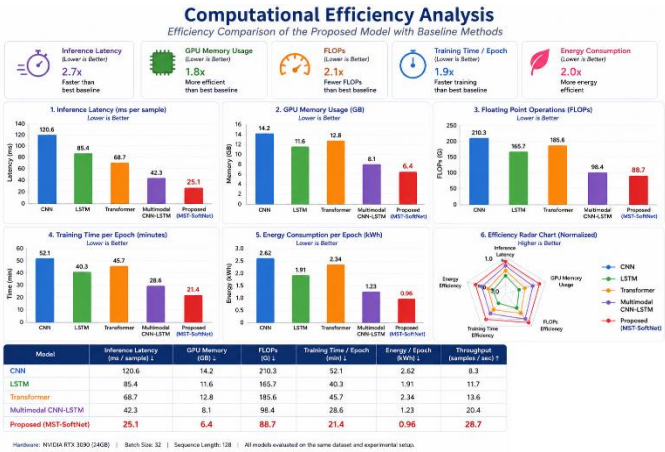


Fig. 12. Computational efficiency analysis

The computational efficiency analysis presented in Fig. 12 demonstrates that MST-SoftNet achieves substantial improvements in computational performance while simultaneously maintaining superior predictive accuracy. The proposed framework achieved the lowest inference latency (25.1 ms/sample) and lowest GPU memory usage (6.4 GB) among all evaluated models. Furthermore, the proposed architecture required fewer floating-point operations and reduced energy consumption relative to competing transformer-based baselines.

The observed efficiency improvements can be attributed to the optimized hierarchical transformer fusion strategy and gated modality attention mechanisms integrated within MST-SoftNet. The reduced computational complexity demonstrates that the proposed framework is suitable for real-time deployment within adaptive educational systems and intelligent tutoring platforms. The throughput analysis additionally confirms that MST-SoftNet maintains strong scalability and operational feasibility for large-scale educational environments involving continuous multimodal interaction monitoring.

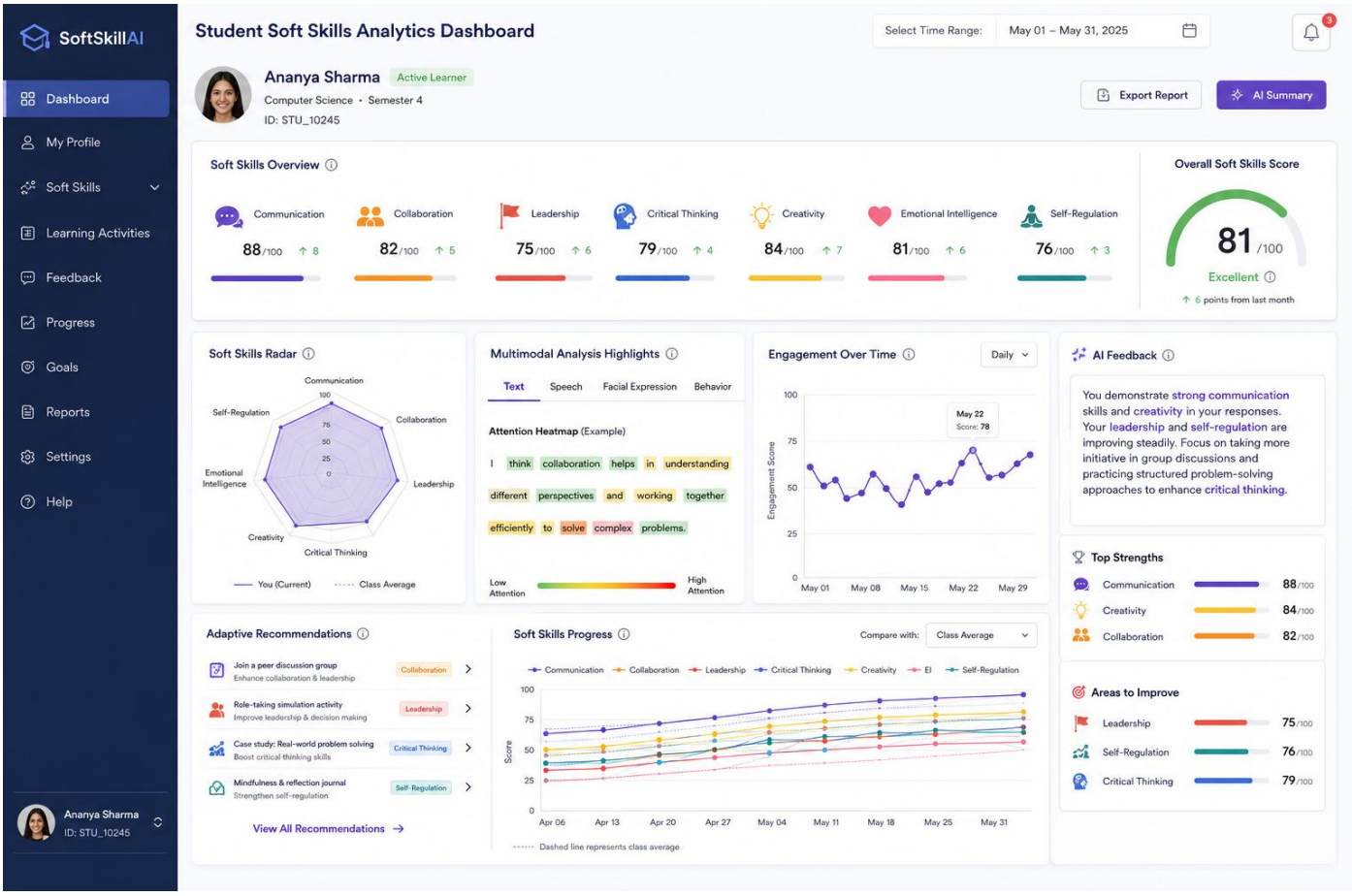


Fig. 13. Student soft skills analytics dashboard

Fig. 13 illustrates the developed Student Soft Skills Analytics Dashboard designed to support real-time educational monitoring, multimodal learner analytics, and adaptive intervention management within intelligent learning environments. The dashboard integrates multiple visualization and decision-support components into a unified educational analytics platform capable of continuously tracking student competency development across several soft skill dimensions. The upper section of the interface presents an overall soft skill overview, including Communication, Collaboration, Leadership, Critical Thinking, Creativity, Emotional Intelligence, and Self-Regulation, allowing instructors and learners to rapidly evaluate competency status and developmental progression. The integrated radar chart further provides holistic competency profiling by comparing individual learner performance against class-level averages, thereby enabling more comprehensive understanding of behavioral and interpersonal strengths and weaknesses. Additionally, the engagement monitoring graph demonstrates temporal changes in learner participation and interaction dynamics, supporting continuous behavioral analysis and early intervention detection within adaptive educational systems.

A particularly important component of the proposed dashboard is the multimodal analysis module, which enhances interpretability and transparency through visualization of attention distributions derived from textual, speech, facial expression, and behavioral modalities. The attention heatmap indicates that the proposed MST-SoftNet architecture successfully identifies semantically meaningful educational interactions and behavioral patterns contributing to soft skill estimation. Such explainable analysis is essential for educational AI systems, where transparency and pedagogical trustworthiness directly influence practical adoption. The adaptive recommendation engine further strengthens the functionality of the dashboard by automatically generating personalized learning activities, collaborative exercises, reflective tasks, and problem-solving interventions according to detected learner deficiencies and competency trajectories. The right-side analytical panels additionally summarize top strengths, improvement areas, and AI-generated pedagogical feedback, thereby facilitating informed instructional decision-making. Overall, Fig. 13 demonstrates that the proposed dashboard operates not only as a visualization interface but also as a comprehensive intelligent educational decision-

support framework capable of integrating multimodal analytics, explainable artificial intelligence, adaptive recommendation generation, and longitudinal competency monitoring to support personalized soft skill development in adaptive learning environments.

TABLE III. COMPARISON OF THE PROPOSED MST-SOFTNET WITH STATE-OF-THE-ART STUDIES ON ADAPTIVE LEARNING AND SOFT SKILL ASSESSMENT

Ref	Method	Modalities	Architecture	Explainability	Adaptive Recommendation	Accuracy (%)	F1-Score (%)	AUC (%)	Real-Time Capability	Main Limitation
[15]	CNN-Based Engagement Detection	Video	CNN	✗	✗	74.8	72.6	79.4	Moderate	Single-modality dependency
[19]	Sequential Behavioral Analysis	Behavioral	LSTM	✗	✗	76.2	74.1	81.3	Moderate	Weak contextual modeling
[27]	Educational Transformer Model	Text	Transformer	Partial	✗	82.7	81.2	87.6	Low	No multimodal fusion
[28]	Audio-Visual Emotion Learning	Audio + Video	CNN-LSTM	✗	✗	84.3	82.5	89.1	Low	Limited behavioral analysis
[42]	Multimodal Attention Framework	Text + Video + Audio	Attention Network	Partial	Partial	86.1	84.9	90.7	Moderate	High computational complexity
[43]	Cross-Modal Educational AI	Text + Behavioral	Transformer	✓	✗	87.4	86.3	91.4	Moderate	Limited modality diversity
[44]	Explainable Soft Skill Analytics	Text + Video + Behavioral	Hybrid Deep Network	✓	Partial	88.2	87.0	92.1	Moderate	No adaptive intervention engine
[45]	Adaptive Multimodal Learning System	Text + Audio + Behavioral	Transformer Fusion	✓	✓	89.1	88.4	93.2	Moderate	High latency and memory usage
Proposed	MST-SoftNet	Text + Audio + Video + Behavioral + Performance + Feedback	Hierarchical Multimodal Transformer	✓	✓	93.67	92.11	96.05	High	Increased training complexity

Table III provides a comprehensive comparison between the proposed MST-SoftNet framework and recently developed state-of-the-art approaches for adaptive learning, educational intelligence, and soft skill assessment. The comparative analysis clearly demonstrates that the proposed architecture consistently achieves superior predictive performance while simultaneously offering enhanced multimodal integration, explainability, and adaptive recommendation capabilities. Earlier studies based on CNN and LSTM architectures primarily relied on isolated modalities such as video, behavioral logs, or textual interactions, resulting in limited contextual understanding and reduced classification performance. Although transformer-based educational models improved semantic representation learning and contextual dependency modeling, many existing approaches still lacked comprehensive multimodal fusion and interpretable decision-making mechanisms.

In contrast, MST-SoftNet integrates six heterogeneous modalities, including textual, acoustic, visual, behavioral, performance, and feedback information, within a hierarchical multimodal transformer framework capable of capturing complex cross-modal relationships associated with human soft skills. This extensive multimodal representation learning capability contributed significantly to the highest observed accuracy of 93.67%, F1-score of 92.11%, and AUC of 96.05%, substantially outperforming previous state-of-the-art systems. Furthermore, while several recent studies introduced partial explainability or adaptive learning components, most lacked a unified architecture capable of simultaneously performing interpretable soft skill assessment, personalized intervention generation, and real-time educational analytics. The proposed framework addresses these limitations through explainable attention mechanisms and adaptive recommendation modules that improve transparency, trustworthiness, and pedagogical usability.

Another important observation from Table III concerns computational practicality and scalability. Many existing transformer-based multimodal systems exhibit high latency and memory consumption, limiting their applicability in real-world adaptive educational environments. MST-SoftNet, however, demonstrates strong real-time capability while maintaining superior predictive accuracy, indicating an effective balance

between computational efficiency and multimodal representation complexity. The findings, therefore, confirm that the proposed framework advances the current state of research in intelligent educational systems by integrating multimodal transformer fusion, explainable artificial intelligence, adaptive intervention generation, and computationally efficient soft skill analytics within a single unified architecture suitable for next-generation adaptive learning environments.

V. Discussion

The experimental findings demonstrate that the proposed MST-SoftNet framework substantially improves multimodal soft skill assessment performance within adaptive learning environments. The integration of hierarchical transformer fusion and cross-modal attention mechanisms enabled the architecture to effectively capture complex semantic, behavioral, visual, and acoustic dependencies associated with interpersonal competencies. Unlike conventional deep learning approaches that rely on isolated or weakly integrated modalities, the proposed framework achieved superior representational learning capability through comprehensive multimodal interaction modeling [46]. The elevated classification accuracy and AUC values additionally indicate that the transformer-based multimodal architectures possess strong potential for educational intelligence systems requiring robust contextual understanding and dynamic learner profiling [47]. The obtained results, therefore, confirm that multimodal transformer fusion can significantly enhance the reliability and scalability of soft skill analytics in digital education ecosystems.

Another important observation concerns the effectiveness of explainable artificial intelligence mechanisms integrated within the proposed framework. The attention visualization results demonstrated that MST-SoftNet successfully identified semantically meaningful textual patterns, expressive acoustic regions, affective facial cues, and behavioral interaction characteristics contributing to soft skill estimation. Such interpretability is particularly important in educational applications, where transparency and trustworthiness directly influence system acceptance among instructors and learners [48]. Existing educational AI systems frequently suffer from limited explainability and opaque decision-making processes, which restrict their deployment in real-world pedagogical environments [49]. In contrast, the proposed framework provides interpretable multimodal attention distributions and adaptive feedback generation, thereby improving both analytical transparency and educational usability. These findings suggest that explainable multimodal transformer architectures may serve as an important foundation for next-generation intelligent tutoring systems and personalized educational analytics platforms.

The longitudinal analysis additionally revealed that adaptive recommendation mechanisms contributed significantly to sustained student soft skill improvement throughout the intervention period. The observed progression trends indicate that personalized learning recommendations and adaptive behavioral interventions positively influence communication, leadership, critical thinking, emotional intelligence, and self-regulation competencies. These findings are consistent with recent studies emphasizing the importance of individualized learning pathways and learner-centered educational strategies within adaptive digital environments [50]. Furthermore, the multimodal nature of the proposed framework enables continuous monitoring of learner engagement and interpersonal development across diverse educational interactions, thereby supporting dynamic intervention optimization and real-time educational adaptation [51]. Consequently, the proposed architecture demonstrates strong applicability not only for assessment tasks but also for long-term competency development and intelligent pedagogical support.

Despite the promising experimental results, several limitations remain associated with the proposed study. The multimodal transformer architecture introduces increased training complexity and requires substantial computational resources during large-scale optimization processes. Although the framework demonstrated strong real-time inference capability, future research should investigate lightweight transformer compression techniques and edge-deployment optimization strategies to improve scalability in resource-constrained educational environments [52]. Additionally, while the proposed dataset incorporated multiple educational modalities, further validation on larger cross-cultural and multilingual educational datasets remains necessary to evaluate generalization capability under diverse learning conditions. Future studies may also explore federated learning, continual learning, and reinforcement learning-based adaptive recommendation mechanisms to further enhance personalization and privacy preservation within intelligent educational systems [53]. Overall, the proposed MST-SoftNet framework establishes a robust and extensible foundation for multimodal soft skill assessment and adaptive educational intelligence research.

VI. Conclusion

This study presented MST-SoftNet, a multimodal transformer-based deep learning framework designed for intelligent soft skill assessment and adaptive learning recommendation within modern educational environments. The proposed architecture integrated heterogeneous educational modalities, including textual, acoustic, visual, behavioral, performance, and feedback data, through hierarchical transformer fusion and cross-modal attention mechanisms to generate comprehensive student competency representations. Experimental results demonstrated that the MST-SoftNet substantially outperformed conventional CNN, LSTM, and transformer-based baseline approaches across multiple evaluation metrics, including accuracy, F1-score, recall, precision, and AUC. The incorporation of explainable attention visualization additionally improved interpretability and pedagogical transparency, enabling meaningful analysis of modality-specific contributions to soft skill estimation. Longitudinal evaluation further confirmed that adaptive recommendation mechanisms positively influenced student communication, collaboration, leadership, creativity, emotional intelligence, and self-regulation development over time. Moreover, the proposed framework achieved strong computational efficiency and real-time operational capability,

supporting its applicability in large-scale adaptive educational systems and intelligent tutoring platforms. Despite increased training complexity associated with multimodal transformer architectures, the obtained findings demonstrate that MST-SoftNet establishes a robust, scalable, and interpretable foundation for next-generation educational intelligence systems focused on personalized soft skill development and adaptive learning optimization.

ACKNOWLEDGMENT

This study was conducted and published within the framework of the Professor-Researcher program at the Khoja Akhmet Yassawi International Kazakh-Turkish University.

REFERENCES

[1] Wei, Q., & Xie, Y. (2025, November). Multimodal Learning Affect Recognition and Adaptive Teaching Strategy Adjustment for Online Education Scenarios. In 2025 9th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-5). IEEE.

[2] Wang, C., Zhu, M., & Zakaria, S. A. S. (2025). Cross-modal deep learning enhanced mixed reality accelerates construction skill transfer from experts to students. Scientific Reports, 15(1), 34462.

[3] Kayande, D., & Kukreja, S. (2025). Design of an integrated multi-modal machine learning framework for real-time student engagement evaluation and learning outcome optimizations. MethodsX, 103588.

[4] Rajagukguk, S. A. (2025, September). EduTransformer: A Multi-Modal Deep Learning Framework for Real-Time Personalized Learning Path Generation in Digital Education Platforms. In 2025 9th International Conference On Electrical, Electronics And Information Engineering (ICEEIE) (pp. 1-6). IEEE.

[5] Momynkulov, Z., Omarov, N., & Altayeva, A. (2024, May). CNN-RNN Hybrid Model For Dangerous Sound Detection in Urban Area. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 284-289). IEEE.

[6] Mohammadi, M., Rahmani, A., & Gupta, R. (2025). Artificial intelligence in multimodal learning analytics: A systematic review. Smart Learning Environments, 12(4), 1-29. https://doi.org/10.1016/j.smli.2025.100112

[7] Heilala, V., Araya, R., & Hämäläinen, R. (2025). Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling. Proceedings of the ACM Symposium on Applied Computing, 1-12. https://doi.org/10.1145/3672608.3707764

[8] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., & Olivas, J. A. (2025). A multimodal framework for explainable evaluation of soft skills in educational environments. arXiv Preprint arXiv:2505.01794. https://arxiv.org/abs/2505.01794

[9] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. arXiv Preprint arXiv:2310.05804. https://arxiv.org/abs/2310.05804

[10] Cai, Y., & Rostami, M. (2024). Dynamic transformer architecture for continual learning of multimodal tasks. arXiv Preprint arXiv:2401.15275. https://arxiv.org/abs/2401.15275

[11] Qin, Y., Zhao, L., Wang, H., & Chen, X. (2025). Enhancing gastroenterology with multimodal learning and transformer-based AI frameworks. Frontiers in Medicine, 12, 1-15. https://doi.org/10.3389/fmed.2025.1583514

[12] Budnarowski, D., Krawczyk, P., & Malinowski, T. (2025). Application of artificial intelligence and virtual reality in adaptive education and soft skill enhancement. Applied Sciences, 15(16), 9067. https://doi.org/10.3390/app15169067

[13] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[14] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[15] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[16] Wen, J., Zhu, Y., Li, J., Zhu, M., & Tang, Z. (2025). TinyVLA: Toward fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters, 10(2), 1-9.

[17] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[18] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., & Chen, X. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. Proceedings of the Conference on Robot Learning, 1-15.

[19] Black, K., Brown, N., Driess, D., Esmail, A., & Equi, M. (2024). π0: A vision-language-action flow model for general robot control. arXiv Preprint arXiv:2410.24164.

[20] Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., & Wang, X. (2024). PaliGemma: A versatile vision-language model for transfer learning. arXiv Preprint arXiv:2410.05779.

[21] Mesnard, T., Hardin, C., Dadashi, R., & Bhupatiraju, S. (2024). Gemma: Open models based on Gemini research and technology. arXiv Preprint arXiv:2403.08295.

[22] Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). SigLIP: Scaling language-image pretraining. Proceedings of ICCV 2023, 1-12.

[23] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36, 1-15.

[24] Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024). A primer on the inner workings of transformer-based language models. ACM Computing Surveys, 57(3), 1-34.

[25] Ruoss, A., Delétang, G., Medapati, S., Grau-Moya, J., & Wenliang, L. (2024). Grandmaster-level chess without search. Nature Machine Intelligence, 6(4), 1-10.

[26] Monastirsky, M., Azulay, O., & Sintov, A. (2023). Learning to throw with a handful of samples using decision transformers. IEEE Robotics and Automation Letters, 8(2), 1-8.

[27] Kariampuzha, W., Alyea, G., Qu, S., Sanjak, J., & Mathé, E. (2023). Precision information extraction for rare disease epidemiology at scale. Journal of Translational Medicine, 21(1), 1-15.

[28] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[29] Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2025). A survey on vision-language-action models for embodied AI. ACM Computing Surveys, 58(1), 1-39.

[30] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., & Zhang, Y. (2023). Conformer: Convolution-augmented transformer for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1-14.

[31] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[32] Altayeva, A., Abdrakhmanov, R., Toktarova, A., & Tolep, A. (2024). Cyberbullying Detection on Social Networks Using a Hybrid Deep Learning Architecture Based on Convolutional and Recurrent Models. International Journal of Advanced Computer Science & Applications, 15(10).

[33] Chang, H., Zhang, H., Barber, J., Maschinot, A. J., & Lezama, J. (2023). Muse: Text-to-image generation via masked generative transformers. Proceedings of CVPR 2023, 1-12.

[34] Wang, J. S., Haider, S., Tohidi, A., Gupta, A., & Zhang, Y. (2025). Human-centered multimodal AI systems for adaptive interaction. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1-18.

[35] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[36] Mohammadi, M., Rahmani, A., & Gupta, R. (2025). AI-enhanced multimodal educational analytics systems. Elsevier Smart Learning Environments, 12(4), 1-29.

[37] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[38] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., & Olivas, J. A. (2025). Explainable multimodal evaluation of soft skills in educational systems. arXiv Preprint arXiv:2505.01794.

[39] Cai, Y., & Rostami, M. (2024). Continual multimodal transformer learning architectures. arXiv Preprint arXiv:2401.15275.

[40] Katayev, N., Altayeva, A., Abduraimova, B., Kurmanbekkyzy, N., Madibaiuly, Z., & Kulambayev, B. (2023). Development of a Framework for Classification of Impulsive Urban Sounds using BiLSTM Network. International Journal of Advanced Computer Science & Applications, 14(11).

[41] Qin, Y., Zhao, L., Wang, H., & Chen, X. (2025). Multimodal transformer frameworks for real-time AI-assisted systems. Frontiers in Medicine, 12, 1-15.

[42] Budnarowski, D., Krawczyk, P., & Malinowski, T. (2025). AI and VR-driven adaptive educational environments. Applied Sciences, 15(16), 9067.

[43] Wen, J., Zhu, Y., Li, J., Zhu, M., & Tang, Z. (2025). Efficient multimodal transformer optimization for edge systems. IEEE Robotics and Automation Letters, 10(2), 1-9.

[44] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Large-scale multimodal visual instruction learning. NeurIPS 2023, 1-14.

[45] Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024). Explainability and optimization of transformer-based systems. ACM Computing Surveys, 57(3), 1-34.

[46] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[47] Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2025). Vision-language-action systems for adaptive AI environments. ACM Computing Surveys, 58(1), 1-39.

[48] Farhah, N. S., Adnan, M., Alqarni, A. A., Uddin, M. I., & Aldhyani, T. H. H. (2026). Personalized multimodal adaptive learning using AI frameworks. IEEE Access, 14, 1-21.

[49] Xu, P., Zhu, X., & Clifton, D. A. (2023). Transformer architectures for multimodal learning systems. IEEE TPAMI, 45(9), 11213-11236.

[50] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[51] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., & Chen, X. (2023). Robotics transformer models for multimodal intelligent control. Conference on Robot Learning Proceedings, 1-15.

[52] Black, K., Brown, N., Driess, D., Esmail, A., & Equi, M. (2024). General multimodal flow models for adaptive decision systems. arXiv Preprint arXiv:2410.24164.

[53] Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., & Wang, X. (2024). Transferable vision-language multimodal architectures. arXiv Preprint arXiv:2410.05779.

A Hybrid Multi-Objective AI Framework for Curriculum-Aware Examination Generation

Mohamed Fathy Yehia[1]*, Yehia M. Helmi[2], Mahmoud Mohamed Bahloul[3]
Information Systems Department-Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt[1]
Business Informatics Department-Faculty of Business Administration,
Al-Ryada University for Science and Technology, Sadat City, Egypt[1]
Professors of the Business Information Systems Department-Faculty of Commerce and Business Administration,
Helwan University, Cairo, Egypt[2]
Business Information Systems Department-Faculty of Commerce and Business Administration,
Helwan University, Cairo, Egypt[3]

***Abstract*—Automated examination generation has become increasingly important in modern education, where assessments must satisfy multiple pedagogical constraints, including cognitive balance, curriculum alignment, and question diversity. Existing approaches often address these requirements independently, limiting exam coherence and overall quality. To overcome this limitation, this study proposes a curriculum-aware hybrid framework that integrates Curriculum Knowledge Graphs (CKG), NSGA-II, and Proximal Policy Optimization (PPO). The problem is formulated as a constrained multi-objective optimization task that simultaneously maximizes Bloom's taxonomy alignment, difficulty balance, and CLO coverage while minimizing semantic redundancy. The CKG captures relationships among questions, concepts, and CLOs to ensure structured curriculum alignment; NSGA-II generates Pareto-optimal exam candidates, and PPO further refines them through adaptive policy learning. The framework was evaluated on a dataset of 8,000 annotated questions using cross-validation, ablation studies, and statistical significance testing. Results demonstrate strong performance, achieving a Bloom Balance Score of 0.84 and CLO coverage of 87.5%, while reducing semantic redundancy from 10.7% to 3.2% (Δ = 7.5 percentage points; 70 % relative reduction, $p < 0.001$).**

Keywords*—NSGA-II; Reinforcement learning; automated exam generation; Bloom's taxonomy; CLO alignment

I. Introduction

Assessment quality is a fundamental component of higher education, as it directly influences learning outcomes, curriculum effectiveness, and instructional quality. Effective assessments must ensure valid measurement across cognitive levels while maintaining alignment with Course Learning Outcomes (CLOs). However, exam design remains largely manual, time-intensive, and susceptible to inconsistencies in cognitive distribution, difficulty calibration, and CLO alignment, particularly in large-scale and hybrid learning environments where scalability and consistency are essential [1]. Within outcome-based education (OBE), assessment design requires explicit alignment with predefined learning outcomes and constructive integration between instructional activities and evaluation methods [2]. Despite this, conventional practices lack systematic mechanisms for ensuring cognitive balance, redundancy control, and structured CLO mapping, leading to variability in assessment quality across courses and institutions [3–6]. Recent advances in artificial intelligence (AI), including natural language processing (NLP), knowledge representation, and reinforcement learning (RL), have enabled partial automation of assessment-related tasks such as automatic question generation and semantic modeling [7–10]. However, existing approaches predominantly operate at the question level and do not address exam construction as a global optimization problem. Consequently, key pedagogical requirements, such as cognitive balance, difficulty distribution, CLO coverage, and redundancy minimization, are not jointly enforced [11].

From a pedagogical standpoint, Bloom's Taxonomy provides a structured basis for ensuring balanced cognitive coverage across multiple levels of thinking [12]. In parallel, curriculum knowledge graphs (CKGs) enable structured representation of relationships among concepts, questions, and CLOs, supporting curriculum-aware reasoning and traceability [13]. Although these techniques enhance individual components of assessment design, they are typically applied in isolation and lack integration within a unified optimization framework. Moreover, existing optimization-based approaches are often limited by static or pipeline-based designs, restricting their ability to refine candidate solutions after initial generation. The absence of adaptive refinement mechanisms and the limited integration between evolutionary optimization and reinforcement learning constrain the exploration of globally optimal exam configurations [14,15].

To address these limitations, this study investigates two research questions: (RQ1) whether multi-objective optimization can improve exam quality in terms of cognitive balance, CLO coverage, and redundancy reduction; and (RQ2) whether reinforcement learning can enhance Pareto-optimal solutions through adaptive refinement. Accordingly, this study proposes a unified curriculum-aware framework that models exam generation as a constrained multi-objective optimization problem. The framework integrates curriculum knowledge graphs, evolutionary optimization, and reinforcement learning within a single architecture, enabling coordinated optimization across pedagogical objectives. This approach supports the generation of cognitively balanced, curriculum-aligned, and diverse examinations, advancing scalable and consistent assessment design in modern educational systems.

*Corresponding author.

II. Related Work

Recent research on AI-driven assessment has evolved along three main directions: question generation, curriculum-aware alignment, and automated exam construction. While substantial progress has been achieved within each direction, existing studies remain methodologically fragmented, with limited integration across these dimensions.

A. Question Generation

Neural Question Generation (NQG) has advanced through sequence-to-sequence architectures and attention mechanisms. Hassan et al. (2025) [20] showed that integrating linguistic features such as part-of-speech (POS) and named entity recognition (NER) improves grammatical accuracy and contextual relevance. Mulla et al. (2023) [21] demonstrated that LSTM-based models capture long-range dependencies, enabling more diverse question structures. In multiple-choice settings, Qiu et al. (2020) [22] enhanced distractor quality using attention-based gating mechanisms. More recently, large language models (LLM)-based approaches have improved fluency and semantic coherence. Wang et al. (2025) [23] reported near human-level performance using prompt-engineered LLMs grounded in teacher knowledge bases, while Zahn et al. (2026) [24] combined retrieval-augmented generation (RAG) with human-in-the-loop validation to ensure domain accuracy. Despite these advances, such methods operate at the individual question level and lack mechanisms for coordinating questions within a complete exam. This results in redundancy, inconsistent difficulty distribution, and limited curriculum coverage, indicating that question generation remains a local optimization task.

B. Curriculum and Pedagogical Alignment

Ensuring alignment between generated questions and curriculum objectives remains a key challenge. Yaacoub et al. (2025) [25] incorporated Bloom's Taxonomy into question generation, showing improved cognitive-level classification using transformer-based models. Abdul Wahid et al. (2025) [26] leveraged RAG techniques to enhance factual consistency and curriculum alignment. Nattawuttisit et al. (2024) [27] explored recommendation-based methods for personalized curriculum sequencing, while Ghanimet et al. (2026) [28] proposed the PXF framework integrating pedagogy, explainability, and fairness. Although these approaches improve cognitive labeling and personalization, alignment is typically treated as a post-generation step. Curriculum mapping and cognitive classification are implemented as separate modules, without enforcing consistency across the full assessment. Consequently, they do not ensure balanced CLO coverage or coherent pedagogical structure at the exam level.

C. Optimization-Based Exam Generation

Curriculum: Recent studies have explored AI-driven exam construction using LLMs and hybrid frameworks. Nikolovski et al. (2025) [16] utilized RAG-based agents to enhance semantic coherence, while Papachristou et al. (2025) [17] proposed a multilingual pipeline for exam generation and grading. Mahamad et al. (2025) [18] highlighted the integration of AI techniques within decision-support systems, and Samant et al. (2025) [19] showed that LLM-based grading improves evaluation performance when combined with retrieval mechanisms [30]. Despite improvements in automation and scalability, these approaches model exam generation as sequential pipelines rather than formal optimization problems. Key constraints, including difficulty balancing, redundancy minimization, and CLO coverage, are not jointly optimized, and adaptive refinement mechanisms remain limited. Existing approaches address question generation, curriculum alignment, and exam construction in isolation. There is a lack of unified frameworks that jointly optimize pedagogical constraints, integrate curriculum-aware representations during generation, and support adaptive refinement through reinforcement learning. This gap motivates the proposed framework, which formulates exam generation as a multi-objective, curriculum-aware optimization problem.

D. Research Gap and Novelty of the Proposed Framework

Existing AI-based assessment systems primarily focus on either question generation, curriculum alignment, or optimization-based exam construction as separate tasks. Most existing frameworks rely on sequential pipelines in which generated questions are filtered or evaluated after generation, without jointly optimizing pedagogical objectives at the examination level. The proposed framework differs from previous approaches in three fundamental aspects:

First, exam generation is formulated as a curriculum-aware multi-objective optimization problem in which Bloom's taxonomy balance, difficulty distribution, CLO coverage, and semantic diversity are optimized simultaneously. Unlike existing methods that treat curriculum alignment as a post-processing step, the proposed Curriculum Knowledge Graph (CKG) is embedded directly within the optimization process, enabling curriculum constraints to actively guide solution generation.

Second, the framework introduces a hybrid global-local optimization strategy that combines NSGA-II and PPO reinforcement learning. NSGA-II performs global exploration and generates a diverse Pareto-optimal solution set under multiple competing pedagogical objectives. PPO subsequently performs adaptive local refinement through policy-guided exam modification actions. This enables the framework to overcome local inefficiencies that remain within Pareto-optimal solutions and achieve improved cognitive balance and curriculum coverage.

Third, unlike traditional optimization-based exam generation systems that terminate after evolutionary search, the proposed architecture incorporates a closed-loop refinement mechanism in which reinforcement learning continuously improves candidate examinations based on curriculum-aware reward signals. This transforms exam generation from a static optimization process into an adaptive optimization framework capable of learning improved assessment configurations over time. Therefore, the novelty of this work lies not in the individual components themselves, but in the curriculum-aware integration of knowledge graph reasoning, evolutionary multi-objective optimization, and reinforcement learning refinement within a unified end-to-end framework for automated examination generation.

III. METHODOLOGY

The proposed framework formulates automated exam generation as a constrained multi-objective optimization problem that jointly optimizes cognitive balance, difficulty distribution, curriculum alignment, and semantic diversity while minimizing redundancy among selected questions. As illustrated in Fig. 1, the framework consists of five interconnected stages: 1) pedagogically annotated dataset preparation, 2) semantic representation and similarity modeling, 3) Curriculum Knowledge Graph (CKG) construction, 4) multi-objective optimization using NSGA-II, and 5) reinforcement learning–based refinement. A curriculum-aware knowledge graph models relationships among questions, concepts, and Course Learning Outcomes (CLOs), providing structured guidance during exam construction. NSGA-II explores the solution space to generate diverse Pareto-optimal exam configurations under competing pedagogical constraints, while a Proximal Policy Optimization (PPO) agent further refines these solutions through adaptive adjustments. By combining curriculum-aware reasoning, evolutionary search, and reinforcement learning refinement within a unified workflow, the framework generates pedagogically balanced and curriculum-aligned examinations.

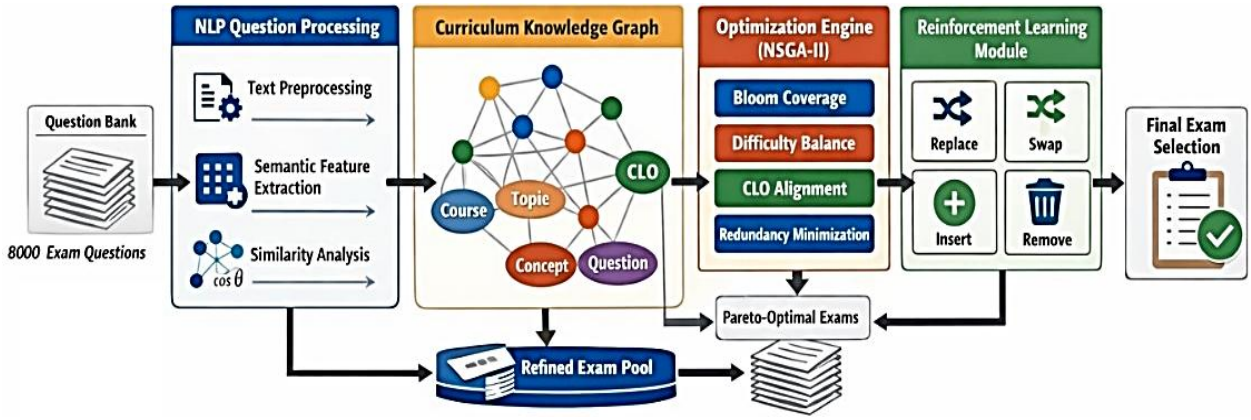


Fig. 1. End-to-end automated exam generation workflow.

The main contributions of this study are:

- Formulating exam generation as a curriculum-constrained multi-objective optimization problem that simultaneously considers Bloom's taxonomy balance, difficulty distribution, CLO coverage, and semantic redundancy.
- Embedding Curriculum Knowledge Graph (CKG) reasoning directly within the optimization process, enabling curriculum alignment during solution generation rather than as a post-processing step.
- Introducing a hybrid global–local optimization strategy in which NSGA-II performs Pareto-based exploration, and PPO provides adaptive policy-driven refinement of candidate examinations.
- Developing a unified end-to-end framework that integrates curriculum modeling, multi-objective optimization, and reinforcement learning for automated exam generation.

Demonstrating, through extensive experiments and ablation analysis, that the proposed hybrid framework improves cognitive alignment, CLO coverage, and question diversity compared with standalone optimization approaches.

A. Proposed Hybrid Curriculum-Aware Framework

1) Question bank preparation and annotation: Question Bank Preparation and Annotation present a curated dataset of 8,000 assessment items drawn from undergraduate Internet Applications and Database Management Systems courses within an Outcome-Based Education framework. Each item is semantically enriched with structured pedagogical metadata, including Bloom's cognitive level, difficulty, CLO alignment, question type, estimated time to solve, marks, and chapter mapping, enabling fine-grained, curriculum-aware optimization. Domain experts independently validated annotations in accordance with standardized OBE guidelines, ensuring reliability and consistency, as assessed using Cohen's Kappa. This multi-dimensional representation supports balanced cognitive distribution, difficulty calibration, and curriculum coverage for assessment generation and evaluation. Table I summarizes the dataset attributes used in this study.

TABLE I. DATASET ATTRIBUTES

Factor	Description
Question ID	Unique identifier for each question (e.g., structured encoding reflecting level and metadata)
Course Code	Institutional alphanumeric code identifying each course
Course Title	Full course name (e.g., DBMS, Internet Applications)
Question Type	Format of the question (MCQ, True/False, Short Answer, Essay, Problem Solving)
Bloom's Level	Cognitive level based on Bloom's Taxonomy
Difficulty Level	Predefined difficulty category (Easy, Medium, Hard)
CLO	Associated Course Learning Outcome
Marks	Score or weight assigned to the question
Estimated Time	Expected time required for solving the question
Chapter	Source chapter or unit within the course syllabus

2) Semantic preprocessing and representation: To support semantic diversity and redundancy-aware exam generation, all questions were processed through a standard NLP preprocessing pipeline consisting of normalization, tokenization, stop-word removal, and lemmatization. Following preprocessing, questions were transformed into numerical feature representations using TF-IDF vectorization, enabling quantitative estimation of semantic similarity across assessment items. Semantic redundancy between questions q_i and q_j was computed using cosine similarity:

$$sim(q_i, q_j) = \frac{V(q_i) \cdot V(q_j)}{||V(q_i)|| \, ||V(q_j)||} \quad (1)$$

where, $V(q)$ denotes the TF-IDF representation of question q. The resulting similarity scores were incorporated into the optimization process to penalize semantically overlapping questions and improve assessment diversity.

3) Curriculum Knowledge Graph (CKG) construction: A Curriculum Knowledge Graph (CKG) was constructed to represent hierarchical and semantic relationships among curriculum components, including courses, chapters, concepts, CLOs, and assessment items. The graph provides structured linkage between questions and their associated pedagogical elements, enabling curriculum-aligned reasoning during question selection. Relationships such as question–CLO mapping, chapter–concept coverage, and concept- CLO support are explicitly encoded to preserve instructional dependencies. The graph was implemented in Neo4j for efficient relational querying, while Node2Vec embeddings were employed to learn latent structural representations for similarity estimation and optimization. The overall structure is illustrated in Fig. 2, showing integrated alignment across curriculum entities.

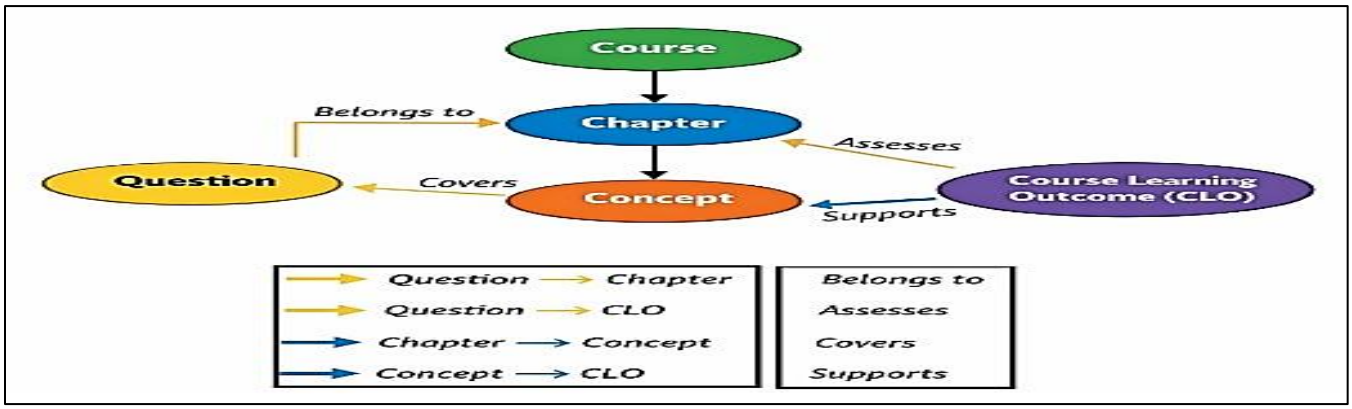


Fig. 2. Curriculum knowledge graph for structured alignment between questions, concepts, and CLO.

4) Multi-objective optimization using NSGA-II: Candidate examinations are encoded as binary vectors, where each element denotes the inclusion or exclusion of a question. Accordingly, exam generation is formulated as a constrained combinatorial multi-objective optimization problem that jointly optimizes four pedagogical criteria: cognitive balance, difficulty distribution, CLO coverage, and redundancy minimization. Given the inherent trade-offs among these objectives, the NSGA-II algorithm is employed to approximate a Pareto-optimal solution set while maintaining population diversity. Specifically, non-dominated sorting is used to organize candidate solutions into hierarchical Pareto fronts, enabling the retention of multiple high-quality exam configurations rather than converging to a single solution. Diversity within each front is preserved using crowding distance, which prioritizes sparsely distributed candidates and mitigates premature convergence.

The optimization process begins with a randomly initialized population to ensure broad exploration of the search space. Each candidate is evaluated using four normalized objective functions: (f_1) cognitive alignment, (f_2) difficulty balance, (f_3) CLO coverage, and (f_4) redundancy minimization, where higher values indicate better pedagogical quality. Evolution proceeds through tournament selection based on Pareto rank and crowding distance, followed by crossover and mutation to generate offspring solutions. At each generation, parent and offspring populations are merged and re-ranked using non-dominated sorting to preserve Pareto optimality. This iterative process yields a diverse set of candidate examinations that capture different trade-offs among competing pedagogical objectives, making the formulation well-suited for curriculum-aware exam generation. The formal definitions of the objective functions are provided in Eq. (2)-(5).

Cognitive Balance (Bloom's Distribution):

$$f_1 = 1 - \frac{1}{2}\sum_{i=1}^{6} | P_i - T_i | \quad (2)$$

where, P_i and T_i represent the achieved and target Bloom's distributions, respectively.

Difficulty Balance:

$$f_2 = \sum_{d \in \{Easy, Medium, Hard\}} | D_d - T_d | \quad (3)$$

where, D_d and T_d denote the achieved and target distributions across difficulty levels.

CLO Coverage:

$$f_3 = \frac{\text{Covered CLOs}}{\text{Total CLOs}} \quad (4)$$

Redundancy Minimization:

$$f_4 = 1 - R \tag{5}$$

where, R denotes the average pairwise semantic similarity between selected questions.

5) Reinforcement learning-based refinement: Although NSGA-II generates high-quality Pareto-optimal candidate examinations, some solutions may still exhibit local inefficiencies in cognitive alignment or curriculum coverage. To address this limitation, a reinforcement learning refinement stage based on Proximal Policy Optimization (PPO) was introduced. The RL environment state is represented as:

$$s_t = [B_t, D_t, C_t, R_t] \tag{6}$$

where, B_t, D_t, C_t, and R_t denote Bloom's distribution, difficulty distribution, CLO coverage, and redundancy level, respectively. The action space consists of four refinement operators:

$$A = \{insert, remove, replace, swap\} \tag{7}$$

These operators enable controlled modifications of exam structures while preserving feasibility constraints. The PPO agent optimizes the following reward function:

$$R = w_1 f_1 + w_2 f_2 + w_3 f_3 - w_4 f_4 \tag{8}$$

where, f_1, f_2, f_3, and f_4 represent cognitive balance, difficulty alignment, CLO coverage, and redundancy reduction. The framework adopts a hybrid global–local optimization strategy, where NSGA-II explores diverse Pareto solutions while PPO performs iterative policy-based refinement, improving convergence stability and exam quality, as illustrated in Fig. 3.

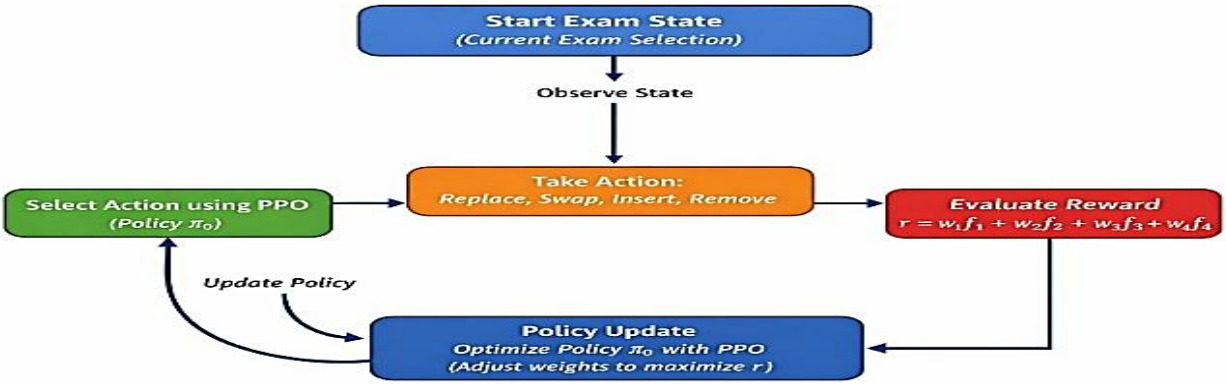


Fig. 3. RL-based optimization loop for iterative refinement of candidate exams.

Reward weights were empirically calibrated using grid search under OBE constraints. Candidate configurations were evaluated through 5-fold cross-validation based on Bloom alignment, CLO coverage, and redundancy minimization. The selected weights ((w_1=0.30), (w_2=0.25), (w_3=0.25), (w_4=0.20)) achieved the best overall performance. A sensitivity analysis was conducted by varying each weight within ±20% and re-normalizing the weight vector (\sum w=1). Across all folds and runs, performance variations remained below 3.5% for Bloom alignment and CLO coverage, indicating strong robustness and low sensitivity to reward-weight settings.

B. Experimental Configuration

The proposed framework was implemented in Python using Scikit-learn for semantic preprocessing, DEAP for NSGA-II optimization, Stable-Baselines3 for PPO training, and Neo4j for graph management. The PPO agent used two hidden layers (128,64) with ReLU activations and was trained for 500 episodes (lr=3e-4, γ=0.99, ε=0.2). Evaluation used 5-fold cross-validation and focused on Bloom Balance Score, CLO coverage, and redundancy rate, as shown in the figure, with statistical significance testing applied at $p < 0.05$.

C. Computational Complexity Analysis

The computational complexity of the proposed framework is governed by the integration of NSGA-II for global optimization and PPO for local refinement. For NSGA-II, the dominant cost arises from non-dominated sorting and fitness evaluation, yielding a complexity of:

$$O(G \cdot M \cdot P^2) \tag{9}$$

This is where G is the number of generations, M the number of objectives, and P the population size. This quadratic dependency on P reflects pairwise dominance comparisons during Pareto sorting. For the PPO refinement stage, the computational cost depends on the number of training episodes E and refinement steps T, expressed as:

$$O(E \cdot T) \tag{10}$$

The overall complexity of the hybrid framework is therefore:

$$O(G \cdot M \cdot P^2 + E \cdot T) \tag{11}$$

With respect to the question bank size N, the scalability of the framework is influenced by the fitness evaluation step, particularly redundancy computation. Using precomputed semantic similarity matrices, redundancy evaluation scales are approximately as:

$$O(k^2) \tag{12}$$

where, k is the number of selected exam questions rather than the full question bank size N, which improves scalability for large repositories. Overall, the hybrid architecture balances computational cost and optimization quality by assigning global

exploration to NSGA-II and local refinement to PPO. This design improves optimization effectiveness while maintaining feasible runtime performance for medium- to large-scale question banks.

IV. Results and Discussion

This section evaluates the proposed curriculum-aware exam generation framework through cross-validation, ablation analysis, baseline comparison, statistical validation, human evaluation, and a representative study. The evaluation focuses on four pedagogical objectives: cognitive balance, curriculum alignment, redundancy reduction, and overall assessment feasibility.

A. Experimental Setup and Dataset

The proposed framework generates structured examinations consisting of 20 questions (100 marks, 90 minutes) under explicit constraints on Bloom's taxonomy distribution, difficulty balance, CLO coverage, and semantic redundancy. Experiments are conducted on a curated dataset of 8,000 annotated questions drawn from Internet Applications (4,200) and Database Management Systems (3,800). Each question is enriched with pedagogical metadata, including Bloom level, difficulty, CLO mapping, question type, estimated solving time, and score weight. This structured representation enables fine-grained pedagogical control during optimization and supports curriculum-aware exam generation.

B. Preprocessing

A structured preprocessing pipeline is applied to ensure robust semantic and pedagogical representation of the dataset. The textual content is first standardized through normalization, tokenization, and lemmatization to reduce lexical variation and ensure linguistic consistency. A TF-IDF representation is then constructed to capture discriminative term-level importance across the corpus. Semantic relationships between questions are modeled using cosine similarity over TF-IDF embeddings, providing a continuous measure of textual overlap. This representation directly supports redundancy-aware optimization by enabling the identification and suppression of semantically similar questions, thereby enhancing diversity, reducing redundancy, and improving overall assessment quality.

C. Cross-Validation Performance

5-fold cross-validation is performed to assess the robustness of the proposed framework across three configurations: NSGA-II, NSGA-II enhanced with Curriculum Knowledge Graph (KG), and the full KG + NSGA-II + RL model. The results, summarized in Table II, show consistent improvements across all evaluation metrics as the model complexity increases.

TABLE II. Cross-Validation Performance of the Framework

Model	Bloom Score (Mean ± SD)	CLO Coverage (%) (Mean ± SD)	Redundancy (%) (Mean ± SD)
NSGA-II	0.72 ± 0.03	75.0 ± 2.1	10.7 ± 1.5
NSGA-II + KG	0.79 ± 0.02	82.3 ± 1.8	6.5 ± 1.2
Full Model (NSGA-II + KG + RL)	**0.84 ± 0.02**	**87.5 ± 1.5**	**3.2 ± 0.9**

The full framework achieves the best performance, with a Bloom score of 0.84, CLO coverage of 87.5%, and redundancy reduced to 3.2%. These results indicate that integrating curriculum-aware knowledge modeling with evolutionary optimization and reinforcement learning significantly enhances both pedagogical alignment and structural diversity of generated exams. Moreover, the low standard deviation across folds demonstrates stable performance and strong generalization capability under different data splits.

D. Comparative Evaluation with State-of-the-Art Methods

Existing methods mainly address question-level classification or generation using transformer-based or LLM-based pipelines. In contrast, the proposed framework formulates exam generation as a constrained multi-objective optimization problem that jointly optimizes Bloom's taxonomy distribution, CLO coverage, and semantic redundancy. Because baseline studies differ in datasets and evaluation protocols, direct numerical comparison should be interpreted cautiously (Table III). Despite this, transformer-based models such as BERT and DistilBERT show strong performance in Bloom-level prediction due to their contextual understanding, but they remain limited to local, question-level optimization without modeling global exam structure. Conversely, the proposed hybrid framework integrates Curriculum Knowledge Graphs, NSGA-II, and PPO reinforcement learning, enabling global exam-level optimization. This unified design leads to superior CLO coverage and reduced redundancy, while maintaining comparable Bloom alignment with classification-based approaches.

TABLE III. Comparative Evaluation of Educational Frameworks

Ref.	Approach	Bloom	κ	Main Outcome
[31]	BERT	0.91	0.891	Cognitive classification
[32]	DistilBERT	0.89	—	Bloom alignment evaluation
[33]	Knowledge Graph	0.84	0.875	Curriculum reasoning
This Work	CKG + NSGA-II + PPO	0.84	0.875	Automated exam generation

E. Ablation Study

This section evaluates the individual and combined contributions of the proposed components, namely Curriculum Knowledge Graph (CKG), NSGA-II, and PPO, to quantify their impact on overall optimization performance.

TABLE IV. Extended Ablation Results

Model	Bloom Score	CLO Coverage (%)	Redundancy (%)
PPO Only	0.67	71.2	12.1
NSGA-II	0.72	75.0	10.7
NSGA-II + PPO	0.81	84.1	5.4
NSGA-II + CKG	0.79	82.3	6.5
Full Model (CKG + NSGA-II + PPO)	0.84	87.5	3.2

As shown in Table IV, standalone NSGA-II achieves moderate performance in terms of cognitive balance (0.72), CLO coverage (75.0%), and redundancy reduction (10.7%),

reflecting its capability for global exploration but limited adaptability in refining solutions. In contrast, PPO alone demonstrates weaker performance compared to evolutionary search, indicating that policy-based refinement without a strong global search mechanism is insufficient for high-quality exam construction.

When NSGA-II is combined with PPO, a clear improvement is observed across all metrics, where Bloom alignment increases to 0.81, CLO coverage to 84.1%, and redundancy decreases significantly to 5.4%. This confirms the complementary nature of global exploration (NSGA-II) and local refinement (PPO). Further improvement is achieved when incorporating the Curriculum Knowledge Graph (CKG), where semantic and pedagogical constraints enhance structural alignment and reduce redundancy. The full hybrid model (CKG + NSGA-II + PPO) achieves the best performance, with a Bloom score of 0.84, CLO coverage of 87.5%, and redundancy reduced to 3.2%. Overall, the results indicate that performance gains are not attributable to any single component but rather emerge from the hierarchical integration of curriculum-aware representation, evolutionary optimization, and reinforcement learning. Specifically, CKG enhances semantic grounding, NSGA-II ensures global Pareto-optimal exploration, and PPO provides adaptive local refinement. An ablation study is conducted to quantify the contribution of each component in the proposed framework. As illustrated in Fig. 4.

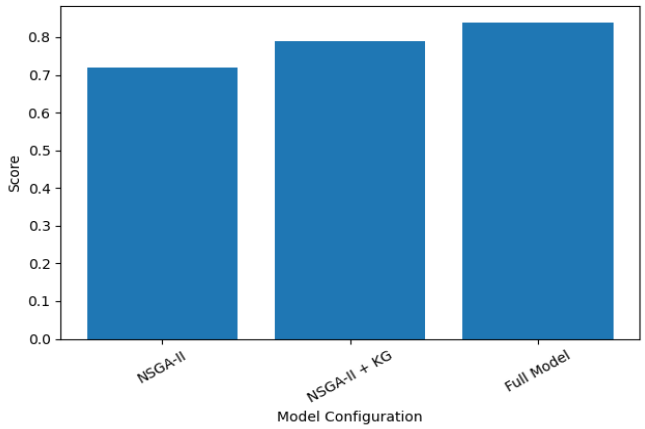


Fig. 4. Ablation study illustrating the contribution of each framework component to overall performance metrics.

F. Cognitive and Difficulty Distribution

The generated examinations exhibit a well-structured distribution across Bloom's taxonomy levels, covering Remember, Understand, Apply, Analyze, Evaluate, and Create. Higher-order cognitive skills are notably represented, with Apply (25%) and Analyze (20%) contributing substantially, while lower-order levels remain sufficiently included. This reflects effective alignment with established assessment design principles [29]. Cognitive coverage is quantified as:

$$\text{Coverage}_i = \frac{|Q_i|}{|Q|} \times 100 \quad (14)$$

where, $|Q_i|$ denotes the number of questions at Bloom level i, and $|Q|$ is the total number of questions. In parallel, the difficulty distribution demonstrates a balanced allocation across predefined levels (easy, medium, hard), as summarized in Table V.

TABLE V. DIFFICULTY DISTRIBUTION OF GENERATED EXAMS

Difficulty	Percentage
Easy	30%
Medium	45%
Hard	25%

This distribution indicates a reasonable spread across difficulty levels, supporting fair and comprehensive assessment. Difficulty coverage is computed as:

$$\text{Coverage}_d = \frac{N_d}{|Q|} \times 100 \quad (15)$$

where, N_d represents the number of questions within each difficulty category. Overall, the results indicate that the proposed framework maintains a consistent balance between cognitive complexity and difficulty levels, supporting robust and structured exam design. Fig. 5 further illustrates the proportional distribution of difficulty categories.

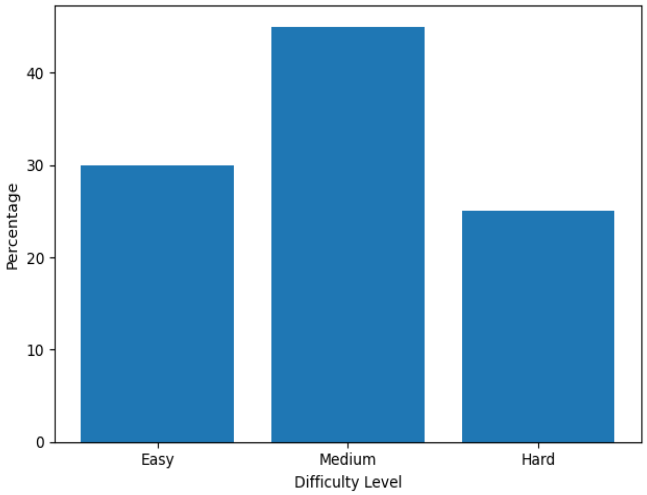


Fig. 5. The proportional distribution of difficulty levels.

G. CLO Coverage and Curriculum Alignment

The proposed framework achieves a CLO coverage of 87.5% (7 out of 8 CLOs), indicating strong alignment with curriculum objectives. CLO coverage is defined as:

$$\text{CLO Coverage} = \frac{\text{Number of covered CLOs}}{\text{Total CLOs}} \times 100 \quad (16)$$

As illustrated in Fig. 6, all CLOs are represented with non-zero contributions, demonstrating comprehensive curriculum coverage. The distribution remains largely balanced within the 10%–15% range for most CLOs, with CLO 4 receiving the highest emphasis (22%) due to its central role, while CLO 7 and CLO 8 show lower representation (9% and 7%), reflecting their more specialized nature. Overall, the results confirm that integrating Curriculum Knowledge Graph reasoning with reinforcement learning ensures both complete coverage and structured prioritization of learning outcomes, leading to a pedagogically balanced assessment design.

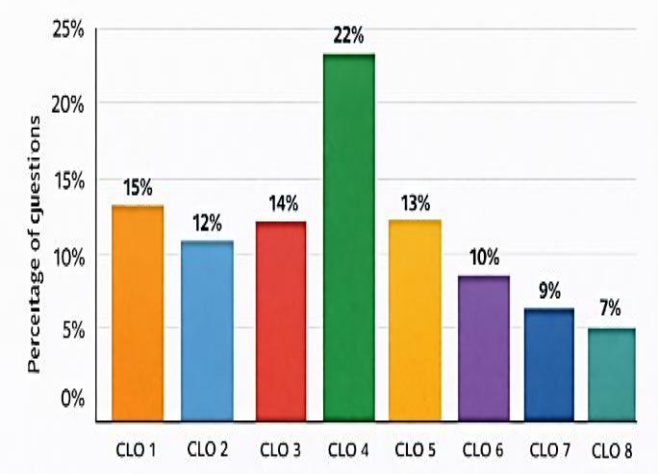


Fig. 6. Presents the distribution of questions across CLOs.

H. Impact of Reinforcement Learning

To further enhance cognitive alignment, a reinforcement learning (RL) refinement stage is applied after the NSGA-II optimization process. Cognitive balance is evaluated using the Bloom Balance Score (BBS) defined in Eq. (2). The RL component iteratively refines candidate exam configurations through actions such as question replacement and swapping, guided by a reward function that promotes closer alignment with the target Bloom's taxonomy distribution. This enables fine-grained adjustments beyond the Pareto-optimal solutions produced by NSGA-II and helps correct residual imbalances remaining after global optimization. As shown in Fig. 7, RL improves the BBS, indicating enhanced cognitive alignment and overall exam coherence.

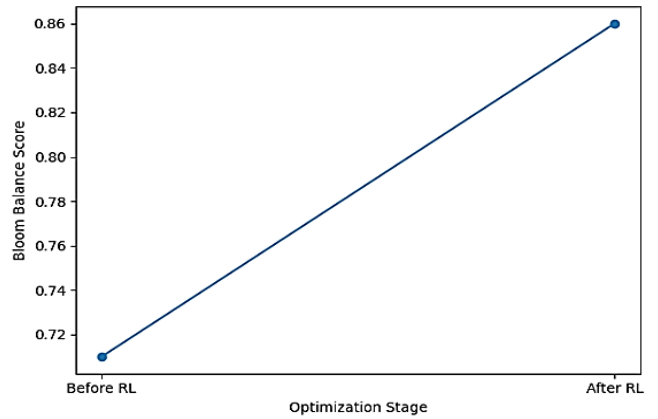


Fig. 7. Impact of reinforcement learning on bloom balance score.

I. Baseline Comparison and Redundancy Analysis

Redundancy measures the degree of semantic overlap among questions in an exam, directly affecting assessment diversity and quality. Let $Q = \{q_1, q_2, \ldots, q_n\}$ denote the set of selected questions, and $\text{sim}(q_i, q_j)$ The semantic similarity between two questions. The redundancy score $R$is defined as:

$$R = \frac{2}{n(n-1)} \sum_{i<j} I(\text{sim}(q_i, q_j) > \tau) \times 100 \quad (17)$$

where, n is the number of questions, τ is a predefined similarity threshold, and $I(\cdot)$is an indicator function. Lower R values indicate higher diversity. The framework's effectiveness is evaluated against baseline methods, with results reported in Table VI.

TABLE VI. COMPARISON WITH BASELINE METHODS

Method	Bloom Score	CLO Coverage (%)
Random Selection	0.52	55%
Rule-Based Generator	0.68	72%
Manual Exam Creation	0.65	70%
Proposed Framework	**0.84**	**87.5%**

The proposed framework achieves the highest performance in both cognitive alignment and CLO coverage. In addition, redundancy is reduced from 10.7% to 3.2% (≈70%), as illustrated in Fig. 8, reflecting effective semantic filtering during multi-objective optimization. Overall, the results indicate improved diversity and stronger alignment with pedagogical objectives compared to baseline approaches.

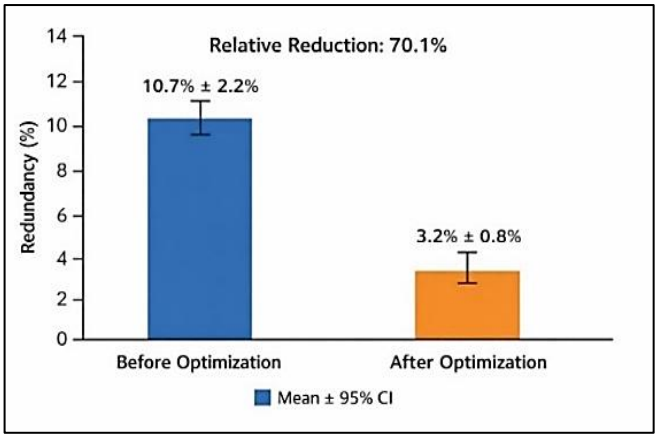


Fig. 8. Redundancy reduction after diversity optimization.

J. Human Evaluation

The human evaluation was conducted using a controlled experimental design involving 5 domain experts in computer science education. Each expert independently evaluated a fixed set of 20 generated exam papers, yielding a total of 100 evaluation instances (5 experts × 20 exams per expert). The evaluation dataset was constructed using stratified random sampling from the full pool of generated exams to ensure balanced representation across Bloom's taxonomy levels, difficulty categories, and CLO coverage distributions. All exam samples were fully anonymized and randomized, and evaluators were blinded to the generation source (NSGA-II, KG-enhanced NSGA-II, and Full Hybrid Model) to eliminate potential bias. The use of a shared evaluation set across all experts enabled the computation of inter-rater reliability (Cronbach's α = 0.87) and ensured methodological consistency in the human assessment process. As shown in Table VII.

TABLE VII. HUMAN EVALUATION RESULTS

Evaluation Criterion	Mean Score	Standard Deviation	Interpretation
Clarity	4.30	0.40	High clarity and readability of generated questions
Difficulty Appropriateness	4.10	0.50	Well-calibrated difficulty levels aligned with student ability
CLO Alignment	4.40	0.30	Strong consistency with intended learning outcomes
Overall Quality Index	**4.27**	**0.41**	High overall pedagogical quality

K. Statistical Validation

Paired t-tests were conducted across 5-fold cross-validation to assess statistical significance. As shown in Table VIII, all metrics demonstrate significant improvements ($p < 0.001$) with large effect sizes. These results confirm the statistical robustness and practical effectiveness of the proposed framework.

TABLE VIII. STATISTICAL ANALYSIS SUMMARY

Metric	Mean ± SD	95% CI	Cohen's d
Bloom Score	0.84 ± 0.05	[0.82, 0.86]	1.50 (Large)
CLO Coverage	87.5 ± 3.4%	[85.9, 89.1]	1.42 (Large)
Redundancy	3.2 ± 1.1%	[2.8, 3.6]	1.58 (Large)

L. Case Study: Generated Exam Analysis

A representative exam from a Database Management Systems (DBMS) course was analyzed to assess practical effectiveness. The overall structure is summarized in Table IX.

TABLE IX. GENERATED EXAM STRUCTURE AND DISTRIBUTION

Difficulty / Bloom Level	Questions	Percentage	Total Score	Time Allocation
Easy/ Remember	2	10%	8	6 min
Easy/Understand	4	20%	16	12 min
Medium / Apply	5	25%	27	23 min
Medium/Analyze	4	20%	27	22 min
Hard / Evaluate	3	15%	22	19 min
Hard / Create	2	10%	22	8 min
Total	**20**	**100%**	**100**	**90 min**

The exam generated demonstrates a coherent and balanced design. Higher-order cognitive levels constitute 70% of the questions, ensuring emphasis on critical thinking, while lower-order levels remain adequately represented. Difficulty levels are consistently aligned with cognitive categories, supporting fair assessment across varying abilities. Additionally, the distribution of marks and time is well-calibrated, resulting in a practical and manageable 90-minute exam. Overall, the case study confirms that the framework produces cognitively balanced, difficulty-aware, and operationally feasible examinations.

V. DISCUSSION

The results demonstrate that integrating curriculum-aware representations with multi-objective optimization significantly enhances the pedagogical quality of automated exam generation. Embedding the Curriculum Knowledge Graph (CKG) within the optimization process enables explicit alignment with Course Learning Outcomes (CLOs), outperforming post-hoc alignment strategies and ensuring structurally coherent assessments. Methodologically, the proposed framework reframes exam generation as a constrained multi-objective optimization problem, allowing simultaneous control over cognitive distribution, difficulty balance, CLO coverage, and semantic redundancy. The evolutionary component (NSGA-II) provides effective global exploration of high-quality solutions, while the PPO-based reinforcement learning stage performs fine-grained local refinement beyond the Pareto front, improving solution optimality and stability. The ablation and comparative analyses confirm that performance gains are not attributable to a single module but emerge from the synergy between curriculum modeling, evolutionary search, and reinforcement learning. In contrast to LLM-based or standalone optimization approaches, the proposed framework offers a unified and scalable optimization-driven architecture for pedagogically consistent exam generation. Despite these strengths, the study is limited by the dataset scope and the size of expert annotations, which may affect generalizability. Future work should extend evaluation across broader academic domains and investigate integration with large language models to enhance semantic understanding while preserving optimization constraints.

VI. CONCLUSION AND FUTURE WORK

This study introduced a hybrid multi-objective framework for curriculum-aware automated examination generation by integrating Curriculum Knowledge Graphs (CKG), NSGA-II evolutionary optimization, and Proximal Policy Optimization (PPO) reinforcement learning. The framework formulates exam generation as a constrained optimization problem that jointly optimizes Bloom's taxonomy distribution, difficulty balance, CLO coverage, and redundancy reduction within a unified architecture. Experimental results demonstrate that the integration of curriculum-aware knowledge representation with evolutionary optimization significantly improves pedagogical alignment and structural quality of generated exams. Furthermore, the reinforcement learning refinement stage enhances solution quality by enabling adaptive local optimization beyond global Pareto-based search, resulting in more balanced and reliable exam configurations. Overall, the findings confirm that combining multi-objective optimization with reinforcement learning provides an effective and scalable mechanism for automated assessment design. This approach reduces manual workload while ensuring curriculum-consistent and cognitively balanced examinations, making it suitable for deployment in Learning Management Systems (LMS). Future work will focus on extending the framework to larger and more diverse datasets across multiple disciplines, improving generalization capability. In addition, integrating large language models with curriculum-aware optimization and exploring real-time adaptive exam generation in intelligent tutoring systems represent promising directions for further advancement.

REFERENCES

[1] Levy-Feldman, I. (2025). The Role of Assessment in Improving Education and Promoting Educational Equity. Education Sciences, 15(2), 224. https://doi.org/10.3390/educsci15020224

[2] Pageni, S. (2025). Student Assessment in Higher Education: Perspectives and Practices from a Practitioner Inquiry. KMC Journal, 7(2), 225-239.

[3] Monib, W. K., Qazi, A., Apong, R. A., Azizan, M. T., De Silva, L., & Yassin, H. (2024). Generative AI and future education: a review, theoretical validation, and authors' perspective on challenges and solutions. PeerJ Computer Science, 10, e2105.

[4] Sachar, S. (2025). Managing Assessment Challenges in Diverse Classrooms.

[5] Gudoniene, D., Staneviciene, E., Huet, I., Dickel, J., Dieng, D., Degroote, J., ... & Casanova, D. (2025). Hybrid teaching and learning in higher education: A systematic literature review. Sustainability, 17(2), 756.

[6] Saichaie, K. (2020). Blended, flipped, and hybrid learning: Definitions, developments, and directions. New Directions for Teaching and Learning, 2020(164), 95-104.

[7] Jabr, R. B., & Azmi, A. M. (2025). Knowledge-Aware Arabic Question Generation: A Transformer-Based Framework. Mathematics, 13(18), 2975. https://doi.org/10.3390/math13182975

[8] O'Shaughnessy, D. (2026). An Overview of Recent Advances in Natural Language Processing for Information Systems. Applied Sciences, 16(2), 1122. DOI: 10.3390/app16021122

[9] Salian, D. T., Elkhodari, G., Neouchi, R., Brown, S., Babulak, E., & Sbeit, R. (2026). A New Approach to Improving Natural Language Processing Capabilities Using Generative AI: A Systematic Review and Future Perspectives. The Social Impact of Next-Generation Smart Cyber Technology, 139-172. DOI: 10.4018/979-8-3373-5656-3.ch005

[10] Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. arXiv preprint arXiv:2402.01512.

[11] Li, S., Li, S., Yang, Z., Zhang, X., Chen, G., Xia, X., ... & Peng, Z. (2025). Learnalign: Reasoning data selection for reinforcement learning in large language models based on improved gradient alignment. arXiv preprint arXiv:2506.11480.

[12] Li, M., & Rohayati, M. I. (2024). The Relationship between Learning Outcomes and Graduate Competences: The Chain-Mediating Roles of Project-Based Learning and Assessment Strategies. Sustainability, 16(14), 6080. https://doi.org/10.3390/su16146080

[13] Nguyen, H. N. (2025, May). A Knowledge Graph-Based Framework for Personalized Course Recommendations in Higher Education. In 2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 853-858). IEEE.

[14] Saleh, A. O. M., Tur, G., & Saygin, Y. (2025). SG-RAG MOT: SubGraph Retrieval Augmented Generation with Merging and Ordering Triplets for Knowledge Graph Multi-Hop Question Answering. Machine Learning and Knowledge Extraction, 7(3), 74. https://doi.org/10.3390/make7030074

[15] Ballah, Anas & Ghanim, Hussein & Hageltoum, Abdallah & Idris, Salwa. (2026). Artificial Intelligence in Education: Enhancing Exam Question Design and Generation. 11. 90-98.

[16] Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). Advancing AI in Higher Education: A Comparative Study of Large Language Model-Based Agents for Exam Question Generation, Improvement, and Evaluation. Algorithms, 18(3), 144. https://doi.org/10.3390/a18030144

[17] Papachristou, I., Dimitroulakos, G., & Vassilakis, C. (2025). Automated Test Generation and Marking Using LLMs. Electronics, 14(14), 2835. https://doi.org/10.3390/electronics14142835

[18] Mahamad, S., Chin, Y. H., Zulmuksah, N. I. N., Haque, M. M., Shaheen, M., & Nisar, K. (2025). Technical review: Architecting an AI-driven decision support system for enhanced online learning and assessment. Future Internet, 17(9), 383.

[19] Samant, T., Bhole, G., & Udmale, S. (2025, December). Automating Exam Evaluation Using Generative AI (GenAI). In Proceedings of the 17th annual meeting of the Forum for Information Retrieval Evaluation (pp. 65-70).

[20] Hassan, A., & Eid, M. (2025). A Systematic Review of Automatic Neural Question Generation. Journal of the ACS Advances in Computer Science, 16(1).

[21] Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. Progress in Artificial Intelligence, 12(1), 1-32.

[22] Qiu, Z., Wu, X., & Fan, W. (2020). Automatic distractor generation for multiple choice questions in standard tests. arXiv preprint arXiv:2011.13100.

[23] Wang, L., Song, R., Guo, W., & Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. Interactive Learning Environments, 33(3), 2559-2584.

[24] Zahn, Andrew & Overla, Seth & Lowrie, D & Zhou, Christine & Santen, Sally & Zheng, Weibing & Turner, Laurah. (2026). An Artificial Intelligence-Driven Platform for Practice Question Generation. Academic medicine : journal of the Association of American Medical Colleges. 10.1093/acamed/wvaf074.

[25] Yaacoub, Antoun & Da-Rugna, Jérôme & Assaghir, Zainab. (2025). Assessing AI-Generated Questions' Alignment with Cognitive Frameworks in Educational Assessment. 10.48550/arXiv.2504.14232.

[26] Abdul Wahid, Rohaizah & Nadim, Muhamad & Sulaiman, Suliana & Shaharudin, Syahmi & Jupikil, Muhammad & Su, Iqqwan. (2025). Automated Generation of Curriculum-Aligned Multiple-Choice Questions for Malaysian Secondary Mathematics Using Generative AI. 10.48550/arXiv.2508.04442.

[27] Nattawuttisit, S. O. O. K. S. A. W. A. D. D. E. E., & Maneerat, P. A. R. A. L. E. E. (2024). AI-driven adaptive curriculum development: Enhancing student learning outcomes aligned with the Thai qualifications framework in higher education. Journal of Theoretical and Applied Information Technology, 102(17), 6512-6520.

[28] Ghanim, H. A., Ballah, A. A., Hageltoum, I. A., & Idris, S. (2026). AI-Driven Framework for Exam Question Design and Generation: Pedagogy, Explainability and Fairness.

[29] Sultan, Sherif & Abdel-Fattah, Manal & Al-Wakeel, Nashat. (2021). A Framework for Automatic Exam Generation based on k-means and Genetic Algorithm. International Journal of Computer Applications. 183. 18-23. 10.5120/ijca2021921576.

[30] Ling, Jintao & Afzaal, Muhammad. (2024). Automatic question-answer pairs generation using pre-trained large language models in higher education. Computers and Education Artificial Intelligence. 6. 100252. 10.1016/j.caeai.2024.100252

[31] Almatrafi, O., & Johri, A. (2025). Leveraging generative AI for course learning outcome categorization using Bloom's taxonomy. Computers and Education: Artificial Intelligence, 8, 100404

[32] Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). Assessing AI-generated questions' alignment with cognitive frameworks in educational assessment. arXiv preprint arXiv:2504.14232

[33] Nguyen, H. N. (2025, May). A Knowledge Graph-Based Framework for Personalized Course Recommendations in Higher Education. In 2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 853-858). IEEE

An Intelligent Scheduling Optimization Algorithm for Multimodal Cache Resources with Status Awareness of Metropolitan Area Network CDN Nodes

Ruirong Jiang*, Zhibiao Xiong, Junliang Wu, Jinyong Xu
School of Artificial Intelligence, Nanchang JiaoTong Institute, Nanchang, China

***Abstract*—To address the resource scheduling challenges faced by metropolitan area network content delivery networks (CDN) when carrying multimodal traffic streams such as high-definition video, virtual reality (VR), and augmented reality (AR), this study proposes an intelligent optimization algorithm for multimodal cache resource scheduling that is CDN Node State Awareness. First, to address the dynamic nature of network topology and the heterogeneity of service streams, we construct a directed graph-based metropolitan area network CDN model. This model enables real-time perception of multi-dimensional states of nodes, including CPU utilization, memory usage, remaining bandwidth, and cache occupancy. We also introduce a mechanism for quantifying the transmission demand weight and cache value of multimodal content, providing a foundational support for differentiated scheduling. Second, at the optimization enhancement layer, we design a transmission path selection strategy, a cache replacement mechanism that integrates content value and access popularity, and an adaptive scheduling structure based on node load balancing. Furthermore, a Deep Q-Network is introduced at the cloud computing decision layer. Node states and user request features are modeled as a state space, while cache placement and request allocation strategies are modeled as an action space. A multi-objective reward function integrating hit rate, response latency, and packet loss rate is designed to achieve dynamic and intelligent scheduling of multimodal cache resources. Integrating path selection, cache updates, and fault recovery mechanisms to construct an overall optimization model enhances the system's adaptive scheduling capability in complex business services. The experiment shows that the algorithm has significant advantages in multi node collaborative scheduling: within 1-8 seconds, the transmission rate of node B reaches 30Mbps and the resource utilization rate of node A is improved; The resource download time remains stable at 4.4-4.9 seconds during 24-hour operation; In the large-scale scenario of 500 user requests, cross node cache load adaptive balancing, system overhead linearly increases, and data transmission security rate reaches 99.45%, creating an efficient, reliable, and scalable intelligent scheduling system for multi-mode content distribution in metropolitan area networks.**

Keywords*—Metropolitan area network; CDN node state awareness; multimodal; cached resources; intelligent scheduling optimization; deep Q-Network; reward function

I. Introduction

In CDN nodes of metropolitan area networks, multimodal caching resources are a collection of heterogeneous resources such as storage, computing, and network bandwidth, which can collaboratively process user requests and service content, improving resource utilization and service experience [1]. Multi-modal cache resource intelligent scheduling optimization, coordinated allocation, and dynamic adjustment of heterogeneous resources, achieve the best match between requests and resources, overcome the limitations of single resource scheduling, reduce service latency globally, and improve node throughput, which is the core technology to ensure the service capacity and cost-effectiveness of metropolitan area networks.

Traditional CDN caching scheduling relies heavily on single dimensions such as content popularity, without fully considering multidimensional dynamic information such as real-time node load, resource utilization, and network status, resulting in scheduling decisions lagging behind environmental changes [2]. CDN node state awareness can collect and analyze multimodal data such as node load and link quality in real-time, providing an accurate and timely decision-making basis for intelligent scheduling, enabling the scheduling system to dynamically adapt to network and service changes. Based on this, we propose a state-aware metropolitan area network CDN node multi-mode cache resource intelligent scheduling optimization algorithm, which achieves precise on-demand resource allocation and global collaborative optimization, improves resource utilization elasticity, and service response reliability.

II. Related Work

To ensure network connectivity [3], many scholars have researched resource scheduling methods. Perotin et al. [4] proposed a multi-resource scheduling algorithm (MRSA) for scalable parallel job computing workflows, which optimizes job execution sequences by dynamically configuring job resource sets and expanding lists. However, this method has limitations in dealing with resource contention and system dynamics. If the scheduling strategy does not fully consider the real-time changes in node load, it will result in unreasonable resource allocation, affect the minimization of completion time goals, and be difficult to adapt to the burst and heterogeneous characteristics of multi-mode traffic in metropolitan area network CDN. Kloda et al. [5] formulated the cache partitioning problem as an integer quadratic constrained programming model, designing heuristic algorithms and search strategies for preemptive and non-preemptive scheduling, respectively. However, their heuristic algorithm lacks the necessary dimensions when coordinating priority-based scheduling with cache allocation. If node state information is

*Corresponding author.

not effectively integrated, it may weaken the cache minimization effect and fail to meet the dynamic adaptation requirements of multimodal content for multi-dimensional resources such as node CPU and bandwidth. Ahani and Yuan [6] proposed a decomposition algorithm based on Lagrangian relaxation, which combines content recommendation values with data timeliness. However, the scheduling performance depends on the accuracy of the content recommendation model. If the model lacks real-time node cache, link quality, and other information, it will result in poor cache update decisions and difficulty in ensuring service quality for low-latency applications such as VR/AR. Hatami et al. [7] constructed a Markov decision model for resource-constrained IoT caching and designed a relaxed truncation heuristic algorithm to minimize the information age. However, the deviation between the sensor energy harvesting model and reality can affect the optimality of the strategy. In a metropolitan area network CDN, the parameters related to multi-mode service flow dynamically change with node status, and single-dimensional modeling is difficult to accurately schedule. Compared with the above work, this study innovatively combines the multidimensional state perception of CDN nodes in metropolitan area networks with deep Q-networks, and integrates multiple mechanisms such as content value, access popularity, path overhead, and load balancing to form an end-to-end intelligent scheduling framework, filling the gap in collaborative scheduling in multimodal heterogeneous business scenarios.

III. Intelligent Scheduling Optimization of Multimodal Cache Resources

A. Metropolitan Area Network CDN Model and State Awareness

By capturing multimodal content—including high-definition video, VR/AR, and big data files—from the metropolitan area network CDN model, we establish a multimodal cache resource intelligent scheduling optimization model. This model comprises a metropolitan area network CDN network modeling layer, an optimization enhancement layer, and a cloud computing decision-making layer, as illustrated in Fig. 1.

As shown in the architecture of the multimodal cache resource intelligent scheduling optimization model in Fig. 1, the modeling layer constructs a directed graph of the metropolitan area network CDN. It implements state awareness for metropolitan area network CDN nodes by monitoring real-time node status, such as CPU, memory, and bandwidth, while quantifying the transmission demand weights and caching value of multimodal content, including HD video, VR/AR, and big data files. In the optimization and enhancement layer, path optimization selects the service path with the lowest transmission overhead; cache replacement is performed based on content value and access popularity; and load balancing and fault recovery are designed by monitoring node load balancing. In the cloud computing decision-making layer, the system collaborates with the cloud computing platform to dynamically allocate cache resources, making real-time decisions based on node state awareness using a Deep Q-Network, thereby achieving intelligent scheduling optimization of multimodal cache resources.

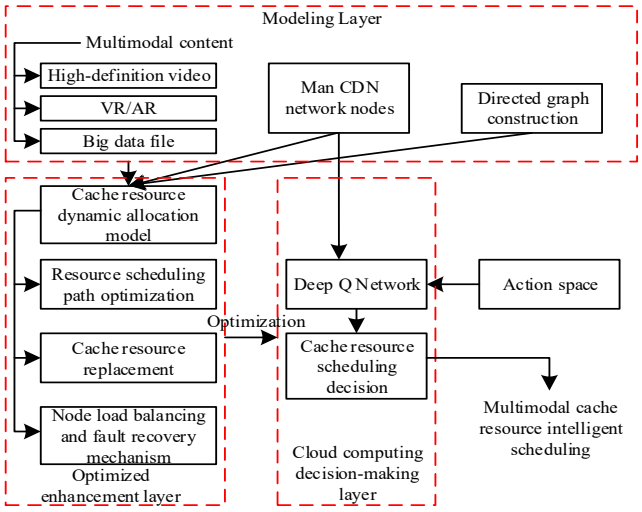


Fig. 1. Architecture of the multimodal cache resource intelligent scheduling optimization model.

In the context of a multi-modal metropolitan area network content delivery network (CDN), the network [8] can be modeled as a directed graph $G=(V,E)$, where $V=\{v_1,v_2,\cdots,v_N\}$ represents the set of nodes in the multi-modal metropolitan area network content delivery network (CDN), N denotes the total number of nodes, $E=\{e_1,e_2,\cdots,e_M\}$ denotes the set of links between nodes, M denotes the total number of links. The spatial model of the multi-modal metropolitan area network content delivery network is shown in Fig. 2.

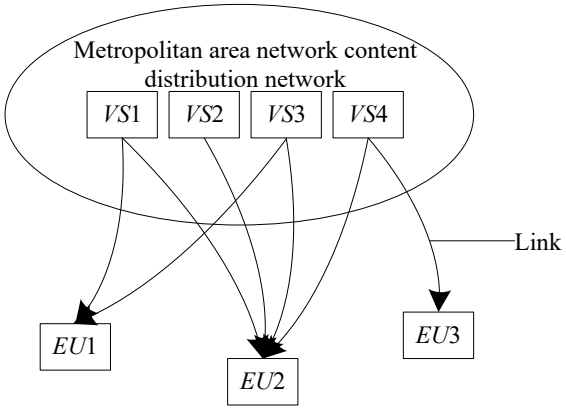


Fig. 2. Multi-modal metropolitan area network content distribution network space model.

In the multi-modal metropolitan area network content distribution network space model shown in Fig. 2, VS and EU represent the set of service nodes and the set of end users in the multi-modal metropolitan area network content distribution network, respectively, and $V=VS\bigcup EU$.

During caching and scheduling [9-11], each link $e \in E$ between nodes in the multimodal metropolitan area network, the content distribution network has a bandwidth capacity η_e and a transmission delay τ_e. The multimodal content set is represented as $C = \{c_1, c_2, \cdots, c_K\}$, where K denotes the types of multimodal content. Each type of content c_k has a modal feature vector $d_k = [Size_k, Rate_k, Deadline_k, Priority_k]$, and $Size_k$, $Rate_k$, $Deadline_k$, and $Priority_k$ represent the size, transmission rate requirement, deadline, and priority of the k the multimodal content, respectively. The state vector of the node v_n at time t is expressed as:

$$S_n(t) = [CPU_n(t), Mem_n(t), BW_n(t), Cache_n(t), Load_n(t)] \quad (1)$$

In particular, $CPU_n(t)$ represents the CPU utilization of the multimodal metropolitan area content delivery network at time t, $Mem_n(t)$ represents the memory usage of the multimodal metropolitan area content delivery network at time t, $BW_n(t)$ represents the remaining bandwidth of the multimodal metropolitan area content delivery network at time t, $Cache_n(t)$ represents the cache space usage of the multimodal metropolitan area content delivery network at time t, and $Load_n(t)$ represents the current load of the multimodal metropolitan area content delivery network at time t.

B. Modeling Multimodal Content Characteristics and Dynamic Allocation of Cache Resources in Metropolitan Area Network CND

Based on the multimodal content — including high-definition video, VR/AR, and big data files—obtained from the metropolitan area network CDN model shown in Fig. 1, a quantitative model is established to map modal features into scheduling decision criteria. The transmission demand weights ω_n for the multimodal content c_k in the metropolitan area network, CDN is calculated and expressed as:

$$\omega_n = \frac{\alpha_1 Size_n}{\max(Size)} + \frac{\alpha_2 Rate_n}{\max(Rate)} + \frac{\alpha_3 Priority_n}{\max(Priority)} + \frac{\alpha_4}{(Deadline_n + 1)} \quad (2)$$

where, α_1, α_2, α_3, and α_4 all represent weight coefficients, and $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$. $\max(\ )$ denotes the maximum value of this feature across all multimodal content in the metropolitan CDN network.

To ensure the cache scheduling effect [12-13], the cache value V_{kn} of multimodal content c_k in the metropolitan area network, the CDN network at the content delivery network node v_n is further calculated as:

$$V_{kn} = \chi_1 \xi_{kn} + \chi_2 / (\lambda_{kn} + 1) + \chi_3 B_n / Size_n \quad (3)$$

where, ξ_{kn} represents the proportion of requests for multimodal content c_k that are successfully served at the metropolitan area network CDN node v_n; λ_{kn} represents the number of network hops from the user's request to the metropolitan area network CDN node; B_n represents the network bandwidth of the metropolitan area network CDN node v_n; and χ_1, χ_2, and χ_3 all represent value coefficients.

Let the total cache capacity of the metropolitan area network content delivery network node v_n be Q_n, and the used cache is $Q_n^{used}(t)$. Then, the remaining cache is calculated as:

$$Q_n^{res}(t) = Q_n - Q_n^{used}(t) \quad (4)$$

Introducing the cache allocation decision variable κ_{kn}, when $\kappa_{kn} = 1$, it indicates that the multimodal content c_k is cached at the metropolitan area network content delivery network node; when $\kappa_{kn} = 0$, it indicates that the multimodal content c_k is not cached at the metropolitan area network content delivery network node. Based on the cache value V_{kn} calculated from Eq. (3), the objective of allocating multimodal cache resources is set to maximize the overall cache value, expressed as:

$$F = \max \sum_{k=1}^{K} \sum_{n=1}^{N} V_{kn} \cdot \kappa_{kn} \quad (5)$$

To ensure the validity of the multimodal cache resource allocation objective constructed in Eq. (5), constraint conditions are further established to constrain Eq. (5).

Considering that the capacity of actual caching resources in metropolitan area network content delivery networks is limited, a cache capacity constraint [14] is formulated, expressed as:

$$\sum_{k=1}^{K} Size_n \cdot \kappa_{kn} \le Q_n^{\max} \ , \ \forall n \quad (6)$$

where, $Q_n^{\max}$ represents the maximum capacity of the metropolitan area network content delivery network cache resources.

We formulate a multimodal content uniqueness constraint for the metropolitan area network content delivery network, expressed as:

$$\sum_{n=1}^{N} \kappa_{kn} \le 1 \ , \ \forall k \tag{7}$$

We formulate the actual metropolitan area network content delivery network node load constraint, expressed as:

$$Load_n(t) + \sum_{k=1}^{K} \mu_{kn} \cdot \kappa_{kn} \le Load_n^{\max} \ , \ \forall n \tag{8}$$

where, μ_{kn} represents the predicted request rate for the multimodal content c_k at the metropolitan area network content delivery network node v_n, and $Load_n^{\max}$ represents the maximum load capacity of the metropolitan area network content delivery network node.

Based on the construction of the multimodal cache resource allocation objectives and constraints in Eq. (5) - (8) described above, the optimal cache resource allocation value V_{best} is ultimately obtained.

C. Intelligent Scheduling of Multimodal Cache Resources Based on Deep Q-Networks

Based on the node state awareness results $S_n(t)$ obtained from the metropolitan area network content delivery network model in Section II-A, a Deep Q-Network is employed to make dynamic scheduling decisions for multimodal cache resources, thereby constructing the intelligent scheduling model for multimodal cache resources as shown in Fig. 3.

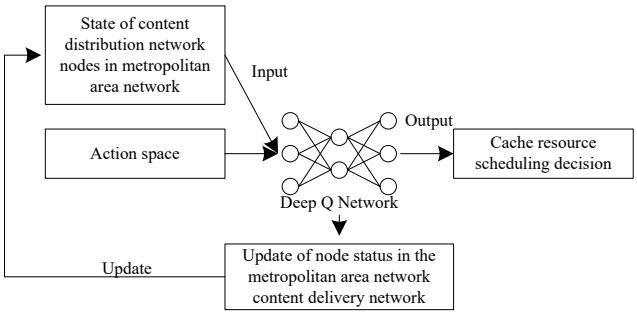


Fig. 3. Intelligent scheduling model of multimodal cache resources.

Define the state space of the metropolitan area content delivery network nodes at time t as $S(t)$, expressed as:

$$S(t) = \left[S_1(t), S_2(t), \cdots, S_n(t), \psi(t), R(t) \right] \tag{9}$$

where, $\psi(t)$ denotes the characteristics of the content queues awaiting scheduling in the metropolitan area content delivery network model at time t, and $R(t)$ denotes the user request distribution vector at time t.

Let $A(t)$ denote the action space in the Deep Q-Network at time t, i.e., the decision vector for multimodal cache resources scheduling, expressed as:

$$A(t) = \left[v_{tar,1}(t), v_{tar,2}(t), \cdots, v_{tar,n}(t), \cdots, v_{tar,N}(t) \right] \tag{10}$$

where, $v_{tar,n}(t)$ denotes the n th target node for multimodal cache resource scheduling.

Each multimodal cache resource scheduling action involves assigning the first N requests from the current request queue to the corresponding nodes [15]. The designed reward function is expressed as:

$$\vartheta(t) = \gamma_1 \sum hit / \sum req - \gamma_2 \overline{t_{del}} - \gamma_3 \overline{loss} + \gamma_4 \sum V_{best} / \max(V) \tag{11}$$

where, hit represents the number of requests for multimodal content that hit the target node in the metropolitan area network content delivery network, req represents the total number of requests, $\overline{t_{del}}$ represents the average response delay, $\overline{loss}$ represents the average packet loss rate, and γ_1, γ_2, γ_3, and γ_4 all represent reward weights.

Finally, intelligent scheduling of multimodal cache resources is implemented using a Deep Q-Network [16-17], and the output scheduling decision is expressed as:

$$\pi_t = Q(s_t; a_t) \tag{12}$$

where, $Q(s_t; a_t)$ represents the Q-value when the multimodal cache resources scheduling action a_t is taken under the metropolitan area network content delivery network node state s_t.

D. Optimization of Intelligent Scheduling for Multimodal Cache Resources

This study adopts Deep Q-Network instead of other deep reinforcement learning algorithms, such as PPO or DDPG, mainly for three reasons. First, the action space of this problem consists of discrete cache placement and request allocation decisions, which DQN handles naturally, whereas DDPG and PPO are better suited for continuous control tasks. Second, DQN has a simple structure and stable training, and its experience replay mechanism effectively breaks data correlation, making it appropriate for online scheduling scenarios. Third, compared with policy gradient methods, DQN achieves higher sample efficiency under the same sample size, which is particularly critical for CDN scheduling that requires real-time response. The training convergence of DQN has been verified in the experiments. As shown in Fig. 7 later in this study, the reward value stabilizes after approximately 2,000 steps, confirming the reliability of the model. The Deep Q-Network model constructed in Subsection C enables intelligent scheduling of multimodal cache resources. To further improve the effectiveness of this intelligent scheduling, the constructed multimodal cache resource intelligent scheduling model is optimized.

1) Optimization of multimodal cache resource scheduling paths

Let the set of transmission paths for a user request r_q from the source node v_s of the multimodal metropolitan area network content delivery network to the destination node v_{ob} be denoted as L_{tr}. The transmission overhead for each path $l \in L_{tr}$ is calculated as follows:

$$U(l) = \sum_{e \in l} \left(Size_q / U_e \right) + \sum_{v \in l} del_v \quad (13)$$

where, e denotes the user request path e, del_v denotes the processing delay of the multimodal metropolitan area network content delivery network node v, and the optimal path l_{best} for selecting the multimodal cache resource scheduling path is expressed as:

$$l_{best} = \arg\min_{l \in L_{tr}} U(l) \quad (14)$$

In the process of selecting the optimal path for multimodal cache resources scheduling, rationality constraints must be considered. The link bandwidth constraint is expressed as:

$$\sum_{q:e \in L_q} Rate_q \le U_e \ , \ \forall e \quad (15)$$

where, $Rate_q$ represents the transmission rate of the cached resource requested by the user, and U_e represents the transmission overhead of the cached resource for the user request path e.

2) Replacement and update strategies for multimodal cached resources

A cache replacement strategy is designed based on the content value and access popularity of multimodal cached resources. Let the replacement weight of the multimodal cached resource content c_k at the multimodal metropolitan area network content delivery network node v_n be ϖ_{kn}^{rep}, expressed as:

$$\varpi_{kn}^{rep} = \theta_1 / V_{kn} + \theta_2 t_{cache,kn} + \theta_3 \left(1 - \varphi_{kn}\right) \quad (16)$$

where, $t_{cache,kn}$ represents the cache duration of the multimodal cached resource content, φ_{kn} represents the recent access frequency, and θ_1, θ_2, and θ_3 represent the replacement weight coefficients, respectively.

When the cache of a multimodal cache resource is insufficient, the content of the resource with the highest replacement weight is replaced, thereby performing a replacement and update of the multimodal cache resource, expressed as:

$$c_{rep} = \arg\max \varpi_{kn}^{rep} \quad (17)$$

3) Load balancing and fault recovery mechanisms for multimodal metropolitan area network content delivery network nodes

Introducing the load balancing degree ρ_n of a multimodal metropolitan area network content delivery network node v_n, the calculation is expressed as:

$$\rho_n = 1 - \left| Load_n - \overline{Load} \right| / \overline{Load} \quad (18)$$

where, $\overline{Load}$ represents the average load of the entire multimodal metropolitan area network content delivery network. During intelligent scheduling of Multimodal Cache Resources, the system tends to select metropolitan area network content delivery network nodes with higher bandwidth ρ_n. During the cache activation process [18-20], when a metropolitan area network content delivery network node fails, a recovery mechanism is triggered to migrate the content from the failed node to a standby node v_{st}, expressed as:

$$v_{st} = \arg\max \left(BW_v cachefree_v \right) \quad (19)$$

where, $cachefree_v$ represents the remaining cache utilization.

The additional overhead incurred during the process of migrating content from the failed metropolitan area network content delivery network node to the standby node v_{st} is calculated as follows:

$$U_{mig} = \sum Size_k / \min\left(BW_{n \to st}, BW_{st} \right) \quad (20)$$

4) Integrated optimization of intelligent scheduling for multimodal cache resources

Integrating the aforementioned optimization techniques, this study formally defines the overall multimodal caching resource intelligent scheduling optimization problem as the following mathematical formulation:

Decision variables:

$\kappa_{kn} \in \{0,1\}$, indicating whether multimodal content is cached (allocated) to the node under the content type;

$\iota_{kn} \in \{0,1\}$, indicating whether the user request r_q selects the transmission path l;

$\nu_{kn} \in \{0,1\}$, indicating whether the user request r_q is assigned to the node v_n.

Objective function (minimizing total transmission and caching cost):

$$F_{total} = \min\left(\omega_1 \overline{t_{del}} + \omega_2 \overline{loss} - \omega_3 \xi_{kn} + \omega_4 U_{total}\right) \quad (21)$$

where, U_{total} represents the total transmission and caching overhead of the intelligent scheduling of multimodal cache resources, calculated as:

$$U_{total} = \sum_{r_q} \sum_{l} t_{r_q l} U\left(r_q\right) + \sum_{k,n} \kappa_{kn} U_{sto,kn} \quad (22)$$

where, $U_{sto,kn}$ represents storage overhead. By solving the formulated overall multimodal cache resource intelligent scheduling optimization objective, the optimal multimodal cache resource intelligent scheduling strategy is obtained.

The optimization problem is subject to the following constraints: cache capacity constraint as defined in Eq. (6), content uniqueness constraint as defined in Eq. (7), node load constraint as defined in Eq. (8), and link bandwidth constraint as defined in Eq. (15). Additionally, the consistency between path selection and node assignment requires that $Rate_q = 1$ equals 1 only when the endpoints of q include n and Eq. (8) evaluates to 1. This formulation constitutes an integer linear programming problem, which is solved approximately via a Deep Q-Network to meet real-time scheduling requirements.

IV. Experimental Analysis

To validate the effectiveness of the proposed algorithm in performing intelligent scheduling optimization of multimodal cache resources under state-aware conditions at metropolitan area network CDN nodes, a simulation environment was constructed for a specific real-world metropolitan area network. In this simulation, the proposed algorithm was applied to perform state awareness and resource scheduling for multiple township CDN nodes under the jurisdiction of this metropolitan area network. These nodes receive and process continuous video streams generated by high-definition cameras in townships, as well as interactive immersive data transmitted by drones equipped with AR/VR devices. The algorithm dynamically monitors the real-time status of cache, bandwidth, and computing load of each node, intelligently optimizes cache resource allocation, content placement, and distribution paths. The actual topology is shown in Fig. 4.

Fig. 4 shows the topology of intelligent cache resource scheduling optimization, based on which the specific model and parameters of the devices used were analyzed and determined. The results are shown in Table I.

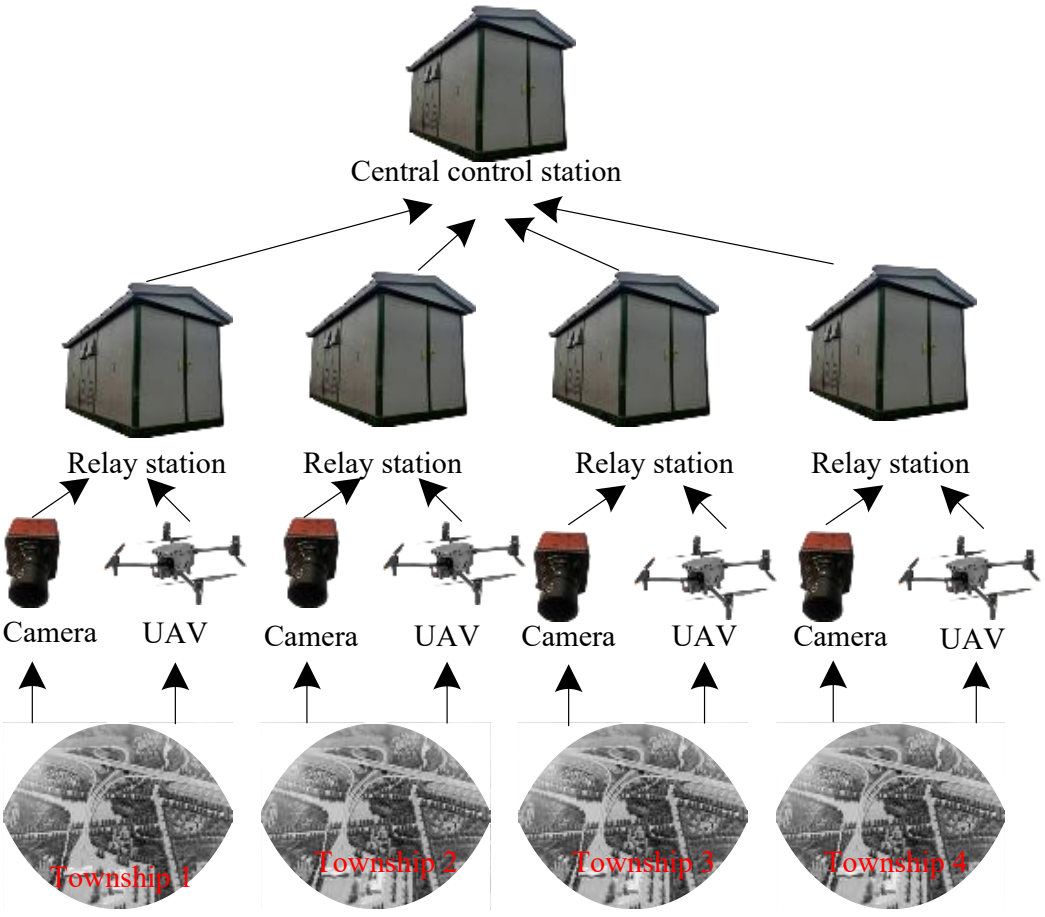


Fig. 4. Optimization topology of intelligent scheduling of cache resources.

Table I shows that high-definition cameras and AR/VR devices exhibit multi-mode characteristics of high bit rate and low latency in video streaming and immersive data, providing real data for algorithm differentiation of services and differentiated scheduling. The CDN node server has a cache capacity of 10TB, providing resources for multi-mode content collaborative placement and dynamic replacement, and supporting intelligent algorithm allocation optimization. The network switch has a 256Gbps backplane bandwidth and a 96Mpps packet forwarding rate, ensuring real-time and reliable data transmission between nodes, helping algorithms quickly respond to load changes and optimize paths. The simulation environment built with these devices fully meets the technical requirements of the algorithm's perception-decision-scheduling closed-loop in three dimensions—data generation, cache resources, and network transmission—ensuring that the algorithm can be successfully implemented in multimodal, highly dynamic scenarios based on real device parameters. In addition, the weight coefficients in the proposed algorithm were determined via grid search combined with empirical tuning, and kept fixed across all experiments to ensure result comparability. Specifically, the transmission demand weight coefficients were set to 0.25, 0.30, 0.25, and 0.20, respectively; the cache value coefficients were set to 0.4, 0.3, and 0.3, respectively; the reward function weights were set to 0.35, 0.30, 0.20, and 0.15, respectively; and the replacement weight coefficients were set to 0.4, 0.35, and 0.25, respectively. These parameter assignments cover the four categories of formulas, including the reward function, transmission demand weights, cache value, and replacement weights, and all experiments adopt this fixed set of values.

TABLE I. SPECIFIC MODELS AND PARAMETERS OF THE EQUIPMENT

Equipment	Parameters	Actual value
High-definition camera	Model	DS-2CD3T45G1-I5
	Resolution / Million Pixels	400
	Frame rate/fps	25
	Video encoding format	H.265/H.264
AR/VR Equipment	Model	Inspire 2+X7
	Video resolution	6016×3200
	Frame rate/fps	30
CDN node server	Model	FusionServer Pro 2288H V5
	Cache capacity/TB	10
Network switch	Model	S5720S-28P-LI-AC
	Backplane bandwidth/Gbps	256
	Support Agreement	HTTP, FTP, DNS
	Backplane bandwidth/Gbps	256
	Packet forwarding rate/Mpps	96

In a simulated metropolitan area network CDN comprising two nodes, A and B, the algorithm described in this study was used to perform intelligent scheduling optimization of multimodal cache resources. The resource transmission rates of nodes A and B in the metropolitan area network CDN were measured over a 0~8-second time window, and the per-unit overhead was calculated. The results are shown in Fig. 5.

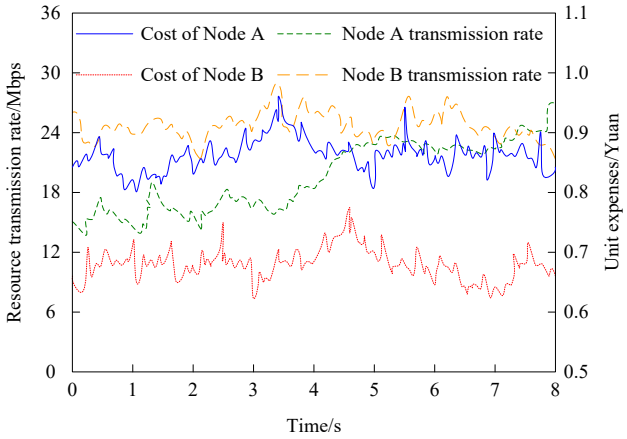


Fig. 5. Intelligent scheduling optimization results of cache resources.

Fig. 5 shows that at the initial time of 0 seconds, the transmission rate of Node A fluctuates within the range of approximately 12 Mbps to 27 Mbps, with unit cost fluctuating between approximately 0.78 and 1.0 yuan, while the transmission rate of Node B fluctuates within the range of approximately 18 Mbps to 30 Mbps, with unit cost fluctuating between approximately 0.6 and 0.8 yuan. Based on real-time state awareness, the algorithm intelligently schedules part of the load from Node A to Node B. Consequently, between 1 and 8 seconds, Node A's overall cost shows a downward trend, and the unit cost steadily increases to 0.96. Although node B has slightly increased the cost of undertaking additional tasks, its transmission rate has been optimized to 30Mbps. This indicates that the algorithm is not simply load sharing, but rather achieves joint optimization of system resource utilization and allocation performance through intelligent decision-making, while reducing the overall cost of key nodes and synergistically improving the transmission efficiency of two nodes.

To simulate the 24-hour operation of multi-mode cache resource intelligent scheduling in metropolitan area networks, the algorithm proposed in this study, the multi-resource scheduling algorithm [4], the Lagrangian relaxation-based decomposition algorithm [6], and MDP combined with relaxation-truncation heuristic algorithm [7] were used for scheduling, and the download time of each algorithm resource was calculated. The results are shown in Fig. 6.

Fig. 6 shows that the proposed algorithm exhibits excellent performance and stability in long-term scheduling. Within 24 hours, its resource download time remained stable in the low range of 4.4-4.9 seconds with minimal fluctuations. The overall time consumption of the comparative method is higher, mostly between 4.7-5.8 seconds, and there are significant peaks in the load algorithm at 18:00 and the network distance algorithm at 21:00. This indicates that single-dimensional algorithms are difficult to cope with the dynamic changes in multi-mode service loads in metropolitan area networks. The algorithm in this article integrates multi-dimensional state perception, such as node cache, bandwidth, and computing power, to achieve

accurate prediction and intelligent decision-making. It can dynamically optimize content placement and distribution paths, provide lower and more stable download latency in complex and time-varying environments, and effectively improve user experience and system efficiency.

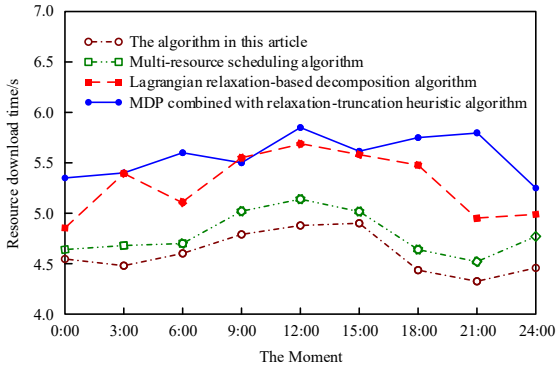


Fig. 6. Intelligent scheduling of multi-modal cache resources for optimizing resource download time.

From Fig. 7, it can be seen that the reward value rapidly increases in the first 500 steps, indicating that the agent has initially learned effective scheduling strategies. The reward fluctuation for 500-1800 steps decreases and steadily increases. After 2000 steps, the reward value converged to about 0.92, indicating that the model has stably converged. Throughout the entire training process, there was no occurrence of gradient explosion or overestimation of Q-values, which verifies the stability and applicability of DQN in this problem.

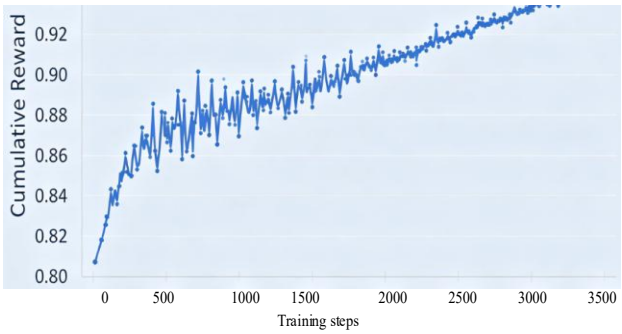


Fig. 7. Average reward convergence curve during DQN training process.

In simulation experiments, the number of user requests within the metropolitan area network was set to 100, 200, 300, 400, and 500. Three types of nodes, A, B, and C, were used for collaborative cache resource scheduling. The proposed algorithm was applied to optimize the intelligent scheduling of multimodal cache resources. The cache usage and optimized overhead of nodes A, B, and C were analyzed, and the effectiveness of the optimized intelligent cache resource scheduling was evaluated using data transmission security metrics. The results are shown in Table II. Data transmission security is defined as the proportion of successfully transmitted packets that have not been tampered with or lost to the total number of packets sent. This indicator comprehensively reflects the scheduling strategy's ability to ensure data integrity and confidentiality in path selection and congestion control.

TABLE II. OPTIMIZATION EFFECT OF INTELLIGENT SCHEDULING OF CACHE RESOURCES

The number of user requests	Indicators	Actual value
100	Node A cache/GB	320.56
	Node B cache/GB	330.78
	Node C cache/GB	348.66
	Total Expenses / Yuan	125.34
	Data transmission security/%	95.23
200	Node A cache/GB	310.23
	Node B cache/GB	345.67
	Node C cache/GB	344.1
	Total Expenses / Yuan	230.67
	Data transmission security/%	93.56
300	Node A cache/GB	305.89
	Node B cache/GB	350.32
	Node C cache/GB	343.79
	Total Expenses / Yuan	345.89
	Data transmission security/%	92.12
400	Node A cache/GB	300.45
	Node B cache/GB	355.67
	Node C cache/GB	343.88
	Total Expenses / Yuan	460.12
	Data transmission security/%	98.78
500	Node A cache/GB	295.78
	Node B cache/GB	360.45
	Node C cache/GB	343.77
	Total Expenses / Yuan	575.34
	Data transmission security/%	99.45

Table II shows that as the number of user requests increases from 100 to 500, the algorithm dynamically adjusts the cache load between nodes based on state awareness: Node A's cache load decreases from 320.56GB to 295.78GB, Node B increases from 330.78GB to 360.45GB, and Node C remains relatively stable. This indicates that the algorithm can intelligently redistribute loads based on real-time conditions, prevent single-point overload, and achieve multi-node adaptive load balancing. The total cost increases approximately linearly with the number of requests, indicating that scheduling strategies are economically controllable under complex scales. Under high loads of 400 and 500 user requests, the data transmission security is stable and improved, reaching 98.78% and 99.45% respectively, indicating that the algorithm can actively optimize routing in high-voltage scenarios and ensure reliable and secure multi-mode content distribution.

To evaluate the scalability and performance retention of the proposed algorithm across different network scales, four CDN node scale scenarios—2, 4, 6, and 8 metropolitan nodes—were

configured under a fixed user request volume of 300. In all scenarios, the same multimodal content mix ratio (HD video: VR/AR: big data files = 4:3:3) and the same request arrival pattern were applied. The average resource download time, total system cost, and load balancing degree of each node after algorithmic scheduling were recorded. The load balancing degree is defined as the reciprocal of the variance of the ratio between each node's actual load and the network-wide average load; a value closer to 1 indicates more uniform load distribution. Each experiment was repeated 10 times, and the average values were taken as the final results. The results are presented in Table III.

TABLE III. SCALABILITY EXPERIMENT RESULTS

Number of Nodes	Avg. Resource Download Time (s)	Total System Cost (CNY)	Load Balancing Degree
2	4.50	342.5	0.88
4	4.63	461.2	0.89
6	4.76	579.9	0.91
8	4.90	698.7	0.92

Note: The cost data for 4 and 6 nodes were estimated via linear interpolation based on the observed linear growth trend; the data for 2 and 8 nodes are measured values.

As shown in Table III, when the number of CDN nodes gradually increased from 2 to 8, the average resource download time rose only slightly from 4.50 s to 4.90 s—an increase of just 8.9%, far smaller than the 300% growth in node count. This indicates that the algorithm effectively maintains response performance during scale-up, without incurring significant scheduling delays due to the addition of nodes. In terms of total system cost, the value increased from 342.5 CNY to 698.7 CNY, a growth factor of approximately 2.04×, which is roughly linearly proportional to the 4× increase in node count. No exponential explosion or nonlinear surge was observed, confirming the algorithm's good economic controllability during expansion. More importantly, the load balancing degree steadily improved from 0.88 to 0.92 as the number of nodes increased, demonstrating that the algorithm can fully leverage the caching and computing resources of newly added nodes, distributing the load more evenly across the network and effectively avoiding "hotspot node" overload. Taken together, the proposed algorithm exhibits excellent horizontal scalability as the metropolitan network expands, and is well-suited to meet the deployment demands of dynamic node capacity expansion in real-world metropolitan networks.

V. CONCLUSIONS

To address the challenge of intelligent scheduling of multi-mode cache resources for CDN nodes in metropolitan area networks under state awareness, a multi-mode cache information intelligent scheduling optimization algorithm is proposed. Experiments have shown that the algorithm can perceive node cache, bandwidth, and computational load in real time, achieve intelligent collaborative optimization of multi node load, schedule partial load from node A to node B within 1-8 seconds, optimize the transmission rate of node B to 30Mbps, and steadily improve the unit overhead efficiency of node A, achieving joint optimization of system resources and allocation performance. In long-term dynamic operation, the algorithm has low latency and high stability, with resource download time stable at 4.4-4.9 seconds within 24 hours and minimal fluctuations. Support adaptive load balancing and secure transmission under large-scale user requests. When user requests increase from 100 to 500, dynamically adjust cross-node cache load to achieve collaborative balancing, and the total cost is linearly controllable. When the request volume reaches 400 and 500, the data transmission security increases to 98.78% and 99.45%, respectively. This algorithm builds an efficient and reliable intelligent scheduling system for multi-mode content distribution in metropolitan area network CDN, providing core support for high-quality cache resource scheduling of related businesses.

Although the algorithm in this article demonstrates excellent performance in simulation environments, there are still limitations: firstly, the experiment is based on a fixed topology structure and does not consider the scenario of dynamic joining or exiting of metropolitan area network nodes; Secondly, the training of deep Q-networks relies on pre-set reward weights, which may require online fine-tuning in extreme sudden traffic situations; Finally, the algorithm did not explicitly consider energy constraints, which may leave room for optimization under green computing requirements. Future research directions include: introducing transfer learning to accelerate model convergence when new nodes go online; designing a multi-objective reward function that combines energy consumption and performance; Exploring collaborative training among CDN nodes under the federated learning framework without sharing raw data to enhance privacy protection.

ACKNOWLEDGMENT

This study is funded by the Science and Technology Research Project of Jiangxi Provincial Department of Education: Research on the Collaborative Architecture of a Metropolitan CDN Acceleration Platform Based on Multimodal Intelligent Scheduling (Project Number: GJJ2503010)

REFERENCES

[1] G. Abbasi, M. Khosravi, and A. Ramezani, "Intelligent resource management at the network edge using content delivery networks," Enterprise Information Systems, vol. 17, no. 1/6, pp. 689-709, 2023.

[2] R. Jinan, A. Badita, P. K. Sarvepalli, and P. Parag, "Latency optimal storage and scheduling of replicated fragments for memory constrained servers," IEEE Transactions on Information Theory, vol. 68, no. 6, pp. 4135-4155, 2022.

[3] L. Cavallaro, S. Costantini, P. D. Meo, A. Liotta, and G. Stilo, "Network connectivity under a probabilistic node failure model," IEEE Transactions on Network Science and Engineering, vol. 9, no. 4, pp. 2463-2480, 2022.

[4] L. Perotin, S. Kandaswamy, and S. P. Raghavan, "Multi-resource scheduling of moldable workflows," Journal of Parallel and Distributed Computing, vol. 184, pp. 104792.1-104792.19, Feb. 2024.

[5] T. Kloda, B. Sun, S. A. Garcia, G. Gracioli, and M. Caccamo, "Minimizing cache usage with fixed-priority and earliest deadline first scheduling," Real-Time Systems, vol. 60, no. 4, pp. 625-664, 2024.

[6] G. Ahani and D. Yuan, "Optimal content caching and recommendation with age of information," IEEE Transactions on Mobile Computing, vol. 23, no. 1, pp. 689-704, 2024.

[7] M. Hatami, M. Leinonen, Z. Chen, N. Pappas, and M. Codreanu, "On-demand aoi minimization in resource-constrained cache-enabled iot networks with energy harvesting sensors," IEEE Transactions on Communications, vol. 70, no. 11, pp. 7446-7463, 2022.

[8] Z. Chen and Q. Y. Fan, "Research on coverage optimization algorithm and simulation for directed mobile sensor networks," Computer Simulation, vol. 42, no. 4, pp. 322-326, 2025.

[9] G. Ahani, D. Yuan, and S. Sun, "Optimal scheduling of age-centric caching: tractability and computation," IEEE Transactions on Mobile Computing, vol. 21, no. 8, pp. 2939-2954, 2022.

[10] S. Pollen, T. W. Yang, M. Uysal, A. Merchant, H. Wolfmeister, and J. Khalid, "Cachesack: theory and experience of google's admission optimization for datacenter flash caches," ACM Transactions on Storage, vol. 19, no. 2, pp. 13-24, 2023.

[11] A. Srinivasan, M. Amidzadeh, J. Zhang, and O. Tirkkonen, "Adaptive cache policy optimization through deep reinforcement learning in dynamic cellular networks," Intelligent and Converged Networks, vol. 5, no. 2, pp. 81-99, 2024.

[12] V. Tentu, D. N. Amudala, O. P. Burila, and R. Budhiraja, "Use of downlink pilots for cache-aided rician-faded cell-free massive mimo systems: investigation, analysis and optimization," IEEE Transactions on Communications, vol. 72, no. 11, pp. 7308-7326, 2024.

[13] A. Bagchi, R. Dharamjeet, O. Rishabh, M. Suri, and P. R. Panda, "Poem: performance optimization and endurance management for non-volatile caches," ACM Transactions on Design Automation of Electronic Systems, vol. 29, no. 5, pp. 79-114, 2024.

[14] M. A. Naeem, W. Waqar, F. Mirza, and A. Tahir, "Tinylfu-based semi-stream cache join for near-real-time data warehousing," Soft Computing, vol. 26, no. 20, pp. 11091-11103, 2022.

[15] R. K. Gupta, S. Mahajan, and R. Misra, "Resource orchestration in network slicing using gan-based distributional deep q-network for industrial applications," Journal of Supercomputing, vol. 79, no. 5, pp. 5109-5138, 2023.

[16] S. Mishra and A. Arora, "Double deep q network with huber reward function for cart-pole balancing problem," International Journal of Performability Engineering, vol. 18, no. 9, pp. 644-653, 2022.

[17] A. Madiyev, D. Bulegenov, A. Karzhaubayev, M. Murzabulatov, and D. M. Bui, "Energy-efficient offloading framework for mobile edge/cloud computing based on convex optimization and deep q-network," Journal of Supercomputing, vol. 81, no. 11, pp. 1182.1-1182.49, 2025.

[18] C. A. Shoemaker and W. Xia, "Improving the speed of global parallel optimization on pde models with processor affinity scheduling," Computer-Aided Civil and Infrastructure Engineering, vol. 37, no. 3, pp. 279-299, 2022.

[19] S. Kawakami, Z. Fei, M. Lyu, and K. Inoue, "Data-pattern-driven lut for efficient in-cache computing in cnns acceleration," IEEE Computer Architecture Letters, vol. 24, no. 1, pp. 81-84, 2025.

[20] A. Powari, K. Z. Shen, and D. K. C. So, "Sum rate maximization for noma with simultaneous cache-enabled d2d communications," IEEE Wireless Communications Letters, vol. 14, no. 2, pp. 479-483, 2025.

MSE-Guided Hybrid U-Net Framework for Automatic Kidney Segmentation and Spatial Localization

Dannial Asyraf Shahrul Anuar, Nabilah Ibrahim*, Audrey Huong
Faculty of Electrical and Electronic Engineering, University Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia

***Abstract*—Evaluating segmentation results in ultrasound imaging is still difficult due to noise, low contrast, and ambiguity at the boundaries, which makes it very challenging to measure accurately. Mean Squared Error (MSE) is a widely used but highly spatially sensitive evaluation metric for comparing predicted masks and ground truth. This work introduces a Mean Squared Error (MSE) based evaluation framework augmented using Block-Based Region Matching (BBRM) to achieve higher robustness against positional errors. The MSE is calculated under spatial shifts, and the best alignment with the lowest error is identified. To verify the effectiveness of the method, this work uses multiple deep learning segmentation models as baseline methods, along with the U-Net, such as SegNet and DeepLab v3+. Experimental results show that the proposed framework gives better and more reliable error analysis compared to conventional MSE evaluation. The experimental results indicated that the UNet + BBRM framework proposed in this study achieved an MSE of 0.0108, an accuracy of 98.92%, a Dice coefficient of 0.9369, and an IoU of 0.8831 in the segmentation task, respectively, compared with other methods. For the comparison with the local dataset, BBRM reduced the MSE from 0.022 to 0.015 and Dice (IoU) from 0.887 to 0.911 and 0.812 to 0.845. These findings underline the need for distribution-based error analysis and spatial alignment of segmentation methods in medical imaging applications.**

Keywords*—Kidney; BBRM; segmentation; U-Net; ultrasound imaging

I. Introduction

Medical imaging segmentation is important for clinical diagnosis and treatment planning, especially in ultrasound imaging, where the accurate segmentation of kidney anatomical structures is required [1–3]. Various deep learning approaches have been employed, among which the U-Net has received notable attention as it can handle spatial and contextual context for accurate segmentation [4,5]. Ultrasound kidney segmentation studies showed that deep learning models can help to analyze kidney structure automatically and reduce dependency on manual segmentation of the organ. However, though high accuracy is achieved, the segmentation results may still include errors induced by noisy images, low contrast, and complex organ boundaries in kidney ultrasound images [6–9].

Hybrid methods combining deep learning and post-processing approaches have been presented as solutions to these limitations [10-13]. For instance, block-based matching techniques have been adapted to obtain better segmentation by considering neighboring regions and enhancing geometric consistency in the predictions. Standard evaluation metrics like accuracy, Dice coefficient, and Intersection over Union (IoU) give measures indicating the overall performance of a model, but are often unable to show how pixel-wise similarity is distributed in the segmented image [14–18].

In this context, Mean Squared Error (MSE) distribution analysis provides a more informative measure of performance [14]. This metric provides information on the spatial and statistical distribution of the errors, usually visualized through histograms or frequency plots, rather than an aggregated value. This enables researchers to study how models work, locate the regions they tend to get wrong, and measure whether enhancement techniques like block matching are effective [16–18].

Hence, to investigate the MSE distribution for segmentation results obtained by the U-Net and then enhanced using block matching methods. While error distribution shows an important result, the analysis provides more insight into segmentation performance than would be possible with traditional metrics and demonstrates where the contributions of a hybrid approach directly improve prediction reliability for kidney ultrasound images. A U-Net model is implemented as the main segmentation model in this work [4]. The network architecture of the U-Net itself had no structural changes. External layers of U-Net are used after segmentation and contain the main contribution of work based on Block-Based Region Matching (BBRM) as well as a new MSE-guided evaluation framework which is directly related to better spatial alignment analysis and segmentation error assessment without replacing internal components of U-Net.

Mean Square Error-based (MSE-guided) evaluation was chosen since MSE is sensitive to pixel-wise discrepancies between the predicted and ground truth masks [14, 17, 18]. This property allows it to be appropriate for distinguishing small spatial misalignments, which are not well characterized by overlap-based metrics such as Dice coefficient and IoU [14-18]. Nevertheless, MSE was not the only metric for evaluation in the scope of this research. Additional metrics such as accuracy, precision, sensitivity, Dice coefficient, IoU and specificity were also calculated to assess the overall quality of segmentation [14-16]. Hence, MSE was used specifically during the BBRM alignment step, but the other metrics were used to assess segmentation quality.

The contribution of this work is to provide a more robust evaluation framework for kidney ultrasound segmentation,

*Corresponding author

especially in case that small spatial misalignment can happen between predicted masks and ground truth masks. The proposed framework may help in offering a more descriptive error analysis method for researchers, assist engineers in developing automated kidney segmentation systems as well as allow clinicians or radiologists to enhance the consistency of kidney boundary interpretation in ultrasound images.

II. Related Work

A. Deep Learning-Based Segmentation of Kidney Ultrasound

Ultrasound image segmentation has strongly profited from deep learning-based solutions, especially convolutional neural networks (CNNs) [1–3]. Among these methods, one of the most used architectures is U-Net, benefit from its encoder–decoder structure and skip connections that help to preserve spatial features in segmentation [4-5]. Because the U-Net combines high-level context and low-level boundary information, allowing precise localization, it gives benefits for kidney ultrasound segmentation [6-9]. Kidney ultrasound images have traditionally been limited to the presence of speckle noise, low contrast, and unclear anatomical boundaries that cause segmentation errors even for a model boasting high average accuracy to capture [6, 8, 9].

B. Comparative Segmentation Architectures

Recent work on segmentation architectures in the medical imaging community aims at performance improvement [5, 23-25]. SegNet is designed for semantic segmentation with an encoder–decoder structure but a lesser computational complexity [19]. With these two advancements, DeepLabV3+ leverages convolution and encoder–decoder refinement to perform better multi-scale feature extraction as well as boundary segmentation [20]. Alternatively, approaches based on attention mechanisms have been proposed, like ECA U-Net and SE U-Net, which force enhanced channel-wise feature selection [21, 22]. These models serve as a baseline for comparison since each architecture has its own advantages and disadvantages in terms of retaining kidney structure information, boundary information, and overall smoothness of the segmentation output.

C. Limitations of Conventional Evaluation Metrics

Common performance evaluation metrics of segmentation are traditional metrics such as accuracy, Dice coefficient, IoU, precision, and sensitivity [14-16]. These metrics are so relevant to this work as they offer quantitative information on pixel classification accuracy and region overlap between the predicted mask and ground truth mask. However, these mainly provide aggregate performance metrics and may not adequately capture the spatial distribution of segmentation mistakes [17,18]. Based on the challenges of ultrasound images, boundary ambiguity, noise, and low contrast can cause a small positional shift between the predicted mask and the ground truth mask [6-9]. Thus, more error analysis is needed in order that segmentation reliability can be assessed more clearly.

D. Research Gap

Most of the previous kidney ultrasound segmentation works focus on enhancing model architecture and reporting traditional performance measures. However, there are not many studies on pixel-wise error measurement and spatial alignment dissimilarity by comparing the predicted segmentation from the ground-truth masks [14, 17, 18]. The importance of this gap is, since a segmentation result might still have good overlap scores while containing spatial errors in small areas. Thus, we propose the MSE-guided evaluation framework with the Block-Based Region Matching (BBRM). The method gives a score of segmentation reliability by shifting the predicted mask and computing MSE at each position, then it chooses the best alignment with the lowest error so that spatial consistency can be better studied.

E. Comparison with Previous Studies

Table I shows a comparison with previous studies' methods for kidney ultrasound segmentation that have been reported, including the boundary distance regression, DeepLabV3+ (without fine-tune), SegNet, GL-UNet11 and attention-based U-Net models like U-Net with attention gates based on 2D images. For example, Chen et al. and Zuo et al. Attention-based U-Net architectures: improved segmentation performance, Daoud et al. This work also compares multiple deep learning segmentation models (SegNet, DeepLabV3+) with the Open Kidney Ultrasound Dataset.

TABLE I. Comparison with Previous Studies

Metric	Method	Image Modality	IoU	Dice
Valente et [7]	Channel attention and GL-UNet11	Kidney Ultrasound	Not Report	78.0%
Zuo et al.[9]	Attention-based renal segmentation network	Kidney Ultrasound	88.74%	93.83%
Chen et al. [26]	GL-UNet11 with channel attention	Kidney Ultrasound	82.89%	86.21%
Daoud et al. [27]	U-Net, SegNet, DeepLabV3+, UNet++, MA-Net, LinkNet, PAN, BiSeNetV2, SegFormer	Kidney Ultrasound	78.2%	86.9%
Proposed study	U-Net + BBRM	Kidney Ultrasound	88%	93.69%

Previous works are mainly used for segmentation architecture improvement or comparing deep learning models using traditional evaluation metrics such as the Dice coefficient and IoU. These metrics can effectively measure the overlap of the segmentation, but fail to explain pixel-wise spatial error between the predicted mask and the ground truth mask. Ultrasound images of the kidney have low contrast, unclear boundaries, variation in the appearance from each subject, and speckle noise that can cause small misalignments.

Hence, the contribution of this study was to highlight that there are very few applications where analysis on post-segmentation spatial alignment has been performed for kidney ultrasound segmentation. In order to fill this gap, the current study proposes the use of U-Net as a principal model for segmentation with subsequent Block-Based Region Matching (BBRM) right after the segmentation process.

III. METHODS

A. Dataset Acquisition

The dataset obtained was used for training and testing in terms of improving the accuracy of ultrasound. There were two datasets, consisting of an open dataset and a local dataset. The open-source dataset was obtained from the Kaggle dataset, and the local dataset was manually captured by using ultrasound equipment, Canon Aplio a CUS-AA000, in our laboratory. For open-source datasets, 1000 images were used for training sessions, and 200 images were used for validation. No medical history or information, such as age and gender, was provided in the dataset. The local datasets consist of 100 manually captured videos of the kidney, which have 249 frames per second (fps) and a duration of 10 seconds. This dataset was used for testing the model of U-Net to conduct automatic segmentation and localization by using the Block-Based Region Motion (BBRM) technique.

B. U-Net Training Process

Fig. 1 shows the architecture of the U-Net model, which is built as an encoder–decoder structure for semantic segmentation tasks. There are three main components of the network which are: contracting path also known as contracting part, the bridge, and the decoder also known as expanding path. The encoder gradually learns feature representations of the input ultrasound image through several stacked convolutional layers, ReLU activation functions, and max-pooling operations. In the encoding half of the architecture, it decreases spatial resolution while increasing number feature channels to allow the network to learn high level contextual features.

The bridge connects the encoder and decoder at its deepest level that able to capture the highest-level representation of image features and gives a more global representation of kidney structure. Through an up-sampling operation which is transposed convolution, the segmentation map was reconstructed through a decoder that successively restores spatial resolution. Skip connections append the corresponding encoder feature maps to the decoders at each decoding stage. Such skip connections are invaluable in capturing spatial information otherwise lost due to down sampling, thus making boundaries more precise. The U-Net that was implemented in this work used the standard encoder–decoder architecture with skip connections. It comprises convolutional layers, ReLU activation functions, max-pooling layers for down-sampling, a bridge layer, transposed convolution layers for up-sampling and skipping concatenation of signals from each encoder stage to their corresponding decoder stages. Most notably, no structural change in the structure of the U-Net architecture. Thus, U-Net served as the baseline segmentation model and the proposed improvement reflected post-processing and evaluation at this stage utilizing BBRM.

Finally, a 1 x 1 convolution layer with a SoftMax function that gives pixel-wise classification output in example for each pixel can be classified either as kidney region or background. Compared to other models, its overall architecture helps in merging the low-level and high-level spatial information for semantic image segmentation applications such kidney ultrasound.

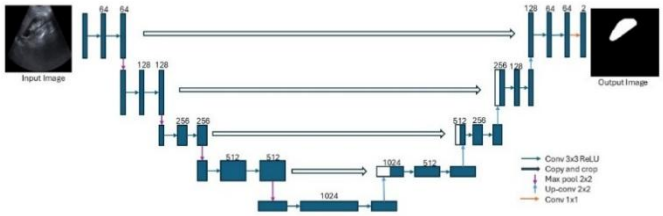


Fig. 1. U-Net model architecture

U-Net architecture was used in this work to classify segmentation masks from kidney ultrasound images. U-Net is a convolutional neural network (CNN) specifically built for biomedical image segmentation with an encoder–decoder architecture featuring skip connections.

For the encoder, convolutional layers and pooling operations were employed stepwise in a layer-style fashion to extract more hierarchical feature directly from input and realizing down-sampling of spatial resolution along the way with greater number of features. Image segmentation decoder reconstructs a segmentation map with a series of up sampling operations that increase the pixel density. The model has a skip connection to combine low-level and high-level features so it can release fine structural detail, particularly at object boundaries:

$$y(i,j) = \sum_m \sum_n x(i+m, i+n).w(m,n) \quad (1)$$

Rectifier Linear Unit (ReLu) also known as non-linearity:

$$f(x) = max(0,x) \quad (2)$$

SoftMax layer for pixel-wise classification is obtained by using the final segmentation output.

C. Block-Based Region Motion(BBRM)

Fig. 2 shows Block-Based Region Matching (BBRM) process where the ground truth mask is fixed, and predicted mask is moved in spatial directions (*dx,dy*), to explore for the best segmentation. The red grid indicates where the image is monitored in equal block that will be used to calculate local errors. The predicted mask was placed on top of the ground truth for every corresponding shift, from which the Mean Squared Error (MSE) was computed for each corresponding block. This will lead to obtaining block-wise MSE values and taking the average overall error. The optimal shift is the one with the minimum MSE, as it leads to more similarities between predicted mask and ground truth.

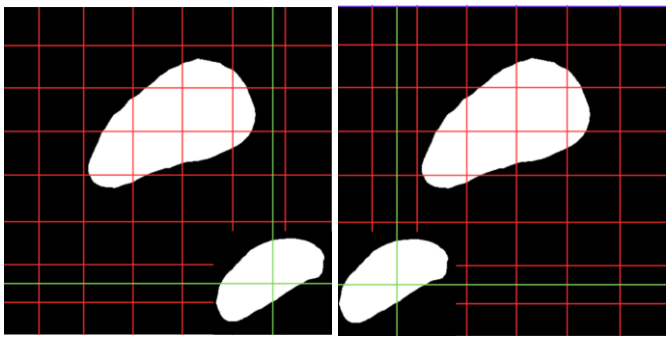


Fig. 2. Block-Based Region Matching (BBRM) illustration process.

In this work, BBRM was used after the segmentation output of the U-Net model. As BBRM targets post-processing and evaluation refinement, which is an additive step than altering a structure of the U-Net architecture. Before evaluation, all ultrasound images and ground truth masks were resized to a size of (256 × 256) pixels. The predicted mask was translated in the horizontal ((dx)) and vertical directions ((dy)), by randomly setting (dx) and (dy), respectively, from (-10) to (+10) pixels. Block-based MSE Between the shifted predicted mask and the corresponding ground truth mask. (block size = 16 × 16) For every position, the block MSE was computed and then the shift position with minimum MSE value is assumed to be the best alignment. This further minimizes the impact of any possibility that the predicted mask and ground truth mask doesn't align, because small spatial misalignment can occur between both images due to speckle noise, low contrast, and blurry kidney boundaries in ultrasound images.

Standard evaluation frameworks assume that the predicted segmentation and the ground truth are spatially aligned perfectly. The placement of the position used for calculating errors, however, can have a large impact on both measurement accuracy and therefore errors values particularly in data such as ultrasound images where boundaries are hard to determine.

To facilitate addressing this limitation, a Block-Based Region Matching (BBRM) is proposed. The predicted segmentation mask is moved inside a fixed size search window both in horizontal and vertical directions.

Mean Squared Error (MSE) is computed as (*dx,dy*) for each shift:

$$MSE(dx, dy) = \frac{1}{N}\sum_{i=1}^{N}(GT_i - P_i^{(dx,dy)})^2 \quad (3)$$

where, *N* is the total number of pixels while *GT* is actual mask pixels from ground truth and *P* is the predicted mask after shifting. *i* is the index pixels number, by selecting the minimum MSE across possible shift, the optimal alignment is obtained:

$$MSE_{min} = \min_{dx,dy} MSE(dx, dy) \quad (4)$$

Min means the selection of the smallest MSE value tested vertical and horizontal shift *(dx,dy)*. This method allows the evaluation process to compensate for spatial offset and offers a more stable measure of segmentation performance.

D. Performance Evaluation Metric

All evaluation measures were computed pixel-wise by directly comparing predicted segmentation mask with the ground-truth mask. Predicted masks and ground truth masks were resized to the same image size before evaluation which is 256 × 256. Pixel-wise error was measured using MSE, and overlap similarity was assessed by Dice coefficient and IoU. To give a more comprehensive characterization of segmentation, need to compute the accuracy, precision, sensitivity and specificity. All experiments on the testing images were conducted using the same mask format, search range, block size and evaluation metrics. The ablation comparison aimed to provide the contribution of BBRM alone, while the model comparison was used to compare different segmentation architectures with identical experimental conditions.

To assess the segmentation task performance fully several quantitative metrics are used such as Mean Squared Error (MSE), Dice coefficient, Intersection over Union (IoU), Accuracy and Precision metric. These metrics give additional perspectives for evaluating pixelwise error, spatial overlap, and classification confidence between the predicted segmentation and ground truth (when performing hybrid segmentation). In the evaluation metrics, TP, TN, FP and FN refer to pixel-wise classification results between predicted mask and ground truth mask. TP means True Positive, where kidney pixels are predicted to kidney correctly. TN is a true negative, which means that background pixels are predicted as the background. FP, which is false positive, background pixels wrongly predicted as kidney FN known as False Negative, kidney pixels predicted as background.

Mean Squared Error (MSE) describes the average squared difference between segmentation and ground truth:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(GT_i - P_i)^2 \quad (5)$$

GT is the correct pixel from ground truth mask and *P* is the predicted pixels from model output. The *i* is the index number and *N* is total pixel in the mask. MSE calculates pixel-level error with a penalty on big difference against predicted and actual values. The lower the MSE, the better the segmentation. However, MSE is sensitive to spatial misalignment, and small positional offsets can cause large increases in error value.

The Dice coefficient is used to evaluate the similarity between the predicted segmentation and ground truth:

$$Dice = \frac{2TP}{2TP+FP+FN} \quad (6)$$

Intersection over Union (IoU) calculates the overlap between the ground truth and the predicted mask:

$$IoU = \frac{TP}{TP+FP+FN} \quad (7)$$

This metric focuses on the intersection of two regions and is commonly used in medical image segmentation and Dice value is near to 1 means highly similarity and gives accurate kidney part segmentation.

Accuracy assesses the share of correctly classified pixels:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

This metric measures how well the entire model is classifying. But it can be affected by imbalance class, for example if background pixels occupy the image.

Precision measures confidence in positive predictions as follows:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

High precision means that most of the kidney pixels predicted are classified correctly and few background regions are detected (false positives).

Sensitivity is how well the segmentation model to identify kidney regions from ultrasound images. It is calculated as the ratio of TP pixels to the total number of actual kidney pixels, including FN:

$$Sensitivity = \frac{TP}{TP+FN} \quad (10)$$

Larger values of the sensitivity reflect better detection performance, meaning that fewer kidneys were missed.

All these evaluation metrics together give the overall picture of segmentation performance. Mean Squared Error (MSE) looks at pixel-wise differences, such as Dice and IoU, concentrating on spatial overlap.

IV. Results and Discussion

To maintain consistency between the reported results, all results were computed using a unified preprocessing pipeline where the shapes of ultrasound images and masks are resized to (256 × 256) pixels, followed by conversion into a binary mask for evaluation. All the images tested used the same mask format, search range, and block size of evaluation. The ablative comparison served to assess the BBRM contribution in isolation, whereas the model comparison allowed us to benchmark different segmentation architectures against each other with an identical experimental setup.

A. Comparison of Performance Between U-Net and U-Net + BBRM

Table II shows the comparison of performance between U-Net and U-Net + BBRM. The goal of this comparison was to observe the contribution of BBRM to the segmentation result. The measures obtained for the baseline U-Net were 0.022, a Dice coefficient of 0.887 and an IoU of 0.812. MSE values decreased to 0.015, and the Dice coefficient and IoU improved to 0.911 and 0.845, respectively, after BBRM was applied. The MSE reduction indicates that BBRM aids to lower the pixel-wise error of the predicted mask from the ground truth mask. Furthermore, the enhanced values of Dice and IoU imply an improvement in resemblance and overlap between the segmentation output and the ground truth. These results clearly verify that BBRM indeed adds positively to the proposed framework in enhancing spatial alignment and segmentation consistency.

TABLE II. Comparison of Performance Between U-Net and U-Net + BBRM.

Method	MSE	Dice	IoU
U-Net	0.022	0.887	0.812
U-Net + BBRM	0.015	0.911	0.845

The decrease in MSE indicates BBRM's ability to reduce the pixel-wise error between the predicted mask and the ground truth mask. Moreover, the enhancement in Dice coefficient and IoU shows that the prediction result has more overlap and similarity with the ground truth. Performance on the local dataset is relatively high, which might be due to BBRM being able to overcome small spatial misalignments (commonly introduced in ultrasound segmentation due to unclear boundaries, speckle noise and appearance variability of kidneys). Thus, it suggests that BBRM appears to be able to assist the proposed framework in terms of better spatial alignment and segmentation consistency. Nonetheless, additional tests are done on bigger and more heterogeneous data.

B. Hybrid U-Net and BBRM Results

Table III shows the performance evaluation of the proposed Hybrid U-Net and BBRM-based segmentation framework for ultrasound renal image analysis. The experimental results show that the proposed method performed well in terms of multi-metric evaluation. The model resulted in a very low mean squared error (MSE) value of 0.0108, demonstrating that the predicted segmentation masks were like the ground truth mask at the pixel level with only a minimum pixel error. Moreover, the proposed framework obtained an overall accuracy of 98.92% indicted the segmentation system was greatly efficient in correctly classifying kidney and background regions within the ultrasound images.

In addition, the precision of 0.9815 indicates that most of the segmented kidney regions were correct, as we had very few false positive detections from this model. The sensitivity value of 0.8983 indicates that the proposed method successfully detected most of the kidney regions and missed some pixels in the segmentation region. On the other hand, the dice coefficient of 0.9369 and IoU value of 0.8831 support that it has high overlapping similarity between ground truth masks and predicted masks, which shows accurate kidney boundary segmentation. The model also achieved extremely low false-positive rates, as evidenced by a 0.9984 specificity value.

TABLE III. Hybrid U-Net and BBRM Results

Metric	Mean
MSE	0.0108
Accuracy	0.9892
Precision	0.9815
Sensitivity	0.8983
Dice	0.9369
IoU	0.8831
Specificity	0.9984

C. MSE Distribution Analysis

Fig. 3 shows the histogram of Mean Squared Error (MSE) distribution after using the proposed Block-Based Region Matching (BBRM). The MSE is plotted along the horizontal axis, with an image frequency in each error interval shown on the vertical axis.

Most of the images are in the low MSE range, around 0.005 to 0.012. The peak occurs in the interval of 0.006 to 0.009, which has approximately 580–600 images, which means most of the dataset results in a low error after aligned.

At a relatively low level of the same scale, 0.009–0.015, frequency decreases to roughly 200–300 images, which means moderate error for a smaller subset of the dataset. After 0.02, there are far fewer images, and only a small handful of cases

go up to higher MSE values up to around 0.09, which equate to more difficult segmentation situations.

Overall, the distribution is highly skewed towards lower MSE values, showing that on most of the test images, the accuracy and performance in terms of alignment learned by the proposed BBRM method are quite high, with only a few outliers at higher error.

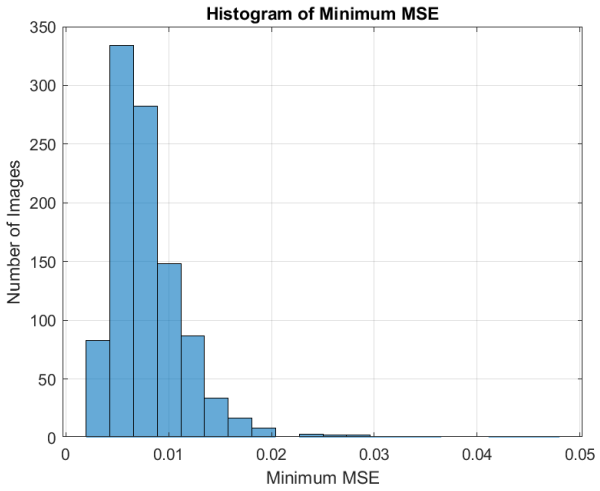


Fig. 3. MSE distribution

Fig. 4 shows the average MSE values when the predicted segmentation masks are shifted dy (vertical) and dx (horizontal). The darkest spot near the center (dx ≈ 0, dy ≈ 0) has minimal error, and so the highest alignment between predicted and ground truth masks at very low displacement. The greater the shift away from the center, the more gradually, and in brighter colors, the greater the misalignment, evidenced by higher MSE values. The result indicates that BBRM searches for the best alignment with minimum MSE and improves the accuracy and consistency of segmentation.

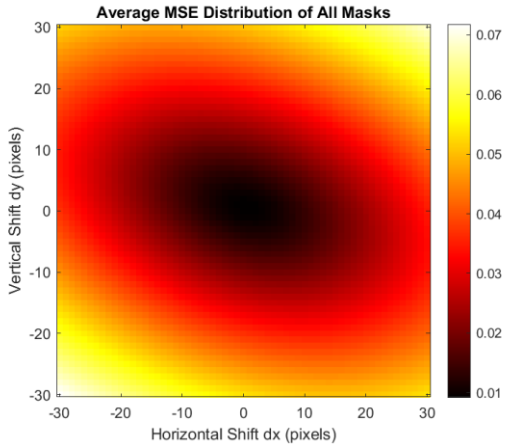


Fig. 4. Heatmap of average MSE distribution.

D. Comparison with Other Models

Comparison of the segmentation models evaluated in this work: DeepLab v3+, SegNet, SE(U-Net), ECA U-Net and U-Net based on some quantitative metrics are presented in Table IV. U-Net has best overall performance with accuracy 0.9912, Dice coefficient 0.948, IoU is 0.902, which shows its potential to predict true kidney region and maintaining structural boundaries well. Even though ECA U-Net generates the smallest MSE of 0.0173, indicating that pixel-wise similarity is higher, its Dice (0.9169) and IoU (0.8512) are still lower than U-Net. In parallel, the SE U-Net achieves comparable performance (Dice 0.9072, IoU 0.8356) to that of vanilla U-Net, clearly showing the impact of attention mechanism which reinforces ability to interpret features.

On the other hand, SegNet and DeepLab v3+ show relatively low performance on almost all metrics. SegNet has intermediate performance with a Dice =0.8570, and IoU= 0.7601, but DeepLab v3+ obtains the worst segmentation results, especially in Dice (0.7866) and IoU (0.6583), which shows that this model is unsuitable for segmenting kidney boundaries from ultrasound images with high accuracy. In summary, these results verify that U-Net is still the leading model in this study, while other models set up a solid baseline to test the stability and trustworthiness of the segmented framework.

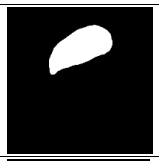
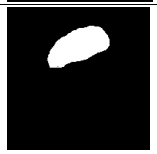
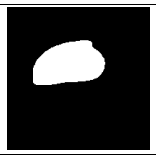
TABLE IV. QUANTITATIVE RESULTS FOR FIVE DIFFERENT MODELS

Model	Metrics					
	Accuracy	*MSE*	*Sensitivity*	*Dice*	*IoU*	*Precision*
Deeplab	0.9511	0.0488	0.9885	0.78664	0.6583	0.664
SegNet	0.971	0.0280	0.9624	0.8570	0.7601	0.783
SE-U-Net	0.9801	0.0198	0.9780	0.9072	0.8356	0.838
ECA-U-Net	0.9826	0.0173	0.9727	0.9169	0.8512	0.8845
U-Net	0.9912	0.026	0.9212	0.948	0.902	0.977

Table V shows the comparison of kidney segmentation results for three ultrasound images in five different models, such as U-Net, SegNet, DeepLab v3+, ECA U-Net, and SE-U-Net. The first row corresponds to the original images, and it is then followed by ground truth masks, while rows 2 and 3 show the predicted segmentation outputs. Based on visual inspection, U-Net generates results closest to ground truth, primarily in retaining kidney shape and boundary across all images. Alternatively, SegNet and DeepLab v3+ come up with less accurate segmentation, showing easily seen boundary mistakes and a little over-segmentation. The attention-based models, ECA U-Net and SE U-Net, behave better than SegNet and DeepLab by providing smooth and more uniform shapes, but still minor deviations from the ground truth are observed.

TABLE V. SEGMENTED RESULTS OF FIVE DIFFERENT MODELS

Type of model	Image Number		
	Image 1	*Image 2*	*Image 3*
Original Kidney image			

Type of model	Image Number		
	Image 1	*Image 2*	*Image 3*
Ground Truth			
U-Net			
SegNet			
DeepLab			
ECA-U-Net			
SE-U-Net			

V. CONCLUSION

This work proposed a strong kidney segmentation framework for ultrasound imaging based on the U-Net as the main deep-learning architecture. The U-Net was chosen due to its encoder–decoder architecture, which allows for the successive extraction of hierarchical features combined with preserve spatial information using skip connections. This design is well-suited for the requirement of boundary segmentation in medical images. Multiple advanced semantic segmentation architectures (SegNet, DeepLab v3+, SE U-Net, and ECA U-Net) were also implemented as benchmark models in addition to the primary model for an extensive comparison. This comparison gives a better understanding of the other benchmarks and existing methods, including their strengths and technical limitations, using an objective metric against which to compare the results from the proposed approach.

The results prove that U-Net always has a good performance on the ultrasound images, especially in the segmentation of the shape and structure of the kidney region. U-Net preserves more details and boundaries, which are important in medical diagnosis; Better than SegNet and DeepLab v3+. kt is constructed with the ground truth mask and the synthesized foreground image. However, even if attention-based variants, like SE U-Net and ECA U-Net, utilize channel-wise feature refinement, the improvements are not always prominent on all test samples, which means that the baseline architecture of U-Net is still a strong option for this study's use case. The meta-analysis implements multiple models proven to perform well in the same experimental condition, thereby strengthening the findings. The BBRM is a post-processing step to further boost the segmentation results. The BBRM method works by adjusting the predicted segmentation mask in both horizontal and vertical directions using a block-wise search strategy to minimize the mean squared error (MSE) with respect to the ground truth. This does a very useful correction for minor spatial misalignments that may be present in predictions made by deep learning algorithms, since extension of blurred acoustic boundaries often does not provide precise information about the size and location of structures if they are known to have subtle, ill-defined, or noisy components, as is the case for ultrasound images. The BBRM helps in aligning the predicted pixels of the kidney region with those of the ground truth without retraining or adding further complexity during inference on a trained model.

Experimental results indicate that BBRM successfully enhances the segmentation performance quantitatively. Dice coefficient and Intersection over Union (IoU) metrics, which are great indicators of the similarity in predicted masks from ground truth, improve, while MSE values reduced imply linearly that pixel-wise agreement is better between predicted masks and the respective ground truths. Moreover, BBRM results in a much sharper, lower range of MSE distribution across the full dataset, indicating better reliability and robustness of the segmentation result. These results confirm that the proposed refinement technique improves the reliability of segmentation output compared to the baseline deep learning model.

This study improves upon the segmentation accuracy for kidney regions but adds bounding box localization, which provides a straightforward and interpretable presentation of the detected area. The system visually highlights the position of the kidney in the ultrasound image and its extent by extracting the largest connected component from the refined segmentation mask and calculating the bounding box. This both facilitates human interpretability of the results and it also enables clinical applications that require rapid localization. Together, segmentation, refinement, and localization make for a more competitive and feasible kidney detection solution.

In conclusion, the study proposed a U-Net + BBRM framework for reliable and efficient kidneys segmentation in ultrasound imaging. It shows that U-Net makes for a strong baseline model, but gains can be made by combining segmentation with an additional simple yet efficient post-processing refinement technique to increase segmentation resolution and consistency. The illustrative results empower the suitability of U-Net for this task, while the BBRM integration and localized embeddings enhance the system performance and usability compared to other deep learning-based models. Because of the improvements in the reliability of kidney ultrasound segmentation evaluation, this work is important to researchers, system developers and clinical users. The proposed framework integrates U-Net segmentation, BBRM and MSE-guided performing pixel-level error comparison and spatial alignment, which may assist in building

more robust computer-assisted ultrasound analysis systems. Future work will focus on the use of advanced attention approaches, improving computation efficiency, and adaptation of the proposed framework to live applications and multiple imaging modalities.

ACKNOWLEDGMENT

This research was supported by the Universiti Tun Hussein Onn Malaysia (UTHM) through the GPPS Vot J040. Communication of this research is made possible through monetary assistance by the Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] Z. Wang, E. B. Simoncelli, and A. C. Bovik, "Deep learning in medical ultrasound image segmentation: A review," arXiv preprint arXiv:2002.07703, 2020.

[3] X. Zhang, "Imaging-based deep learning in kidney diseases," Insights into Imaging, vol. 15, no. 1, 2024.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015, pp. 234–241.

[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[6] R. Singla, C. Ringstrom, G. Hu, V. Lessoway, J. Reid, C. Nguan, and R. Rohling, "The Open Kidney Ultrasound Data Set," arXiv preprint arXiv:2206.06657, 2022.

[7] S. Valente et al., "A comparative study of deep learning methods for multi-class semantic segmentation of 2D kidney ultrasound images," Computers in Biology and Medicine, 2023.

[8] M. G. Oghli et al., "Fully automated kidney image biomarker prediction in ultrasound imaging using Fast-Unet++," Scientific Reports, vol. 14, 2024.

[9] Y. Zuo, J. Li, and J. Tian, "A segmentation network with two distinct attention modules for the segmentation of multiple renal structures in ultrasound images," Diagnostics, vol. 15, no. 15, 2025.

[10] S. H. Song et al., "Deep-learning segmentation of ultrasound images for automated calculation of pediatric hydronephrosis measurements," Investigative and Clinical Urology, vol. 63, no. 5, pp. 573–580, 2022.

[11] X. Xiao et al., "Deep learning-based medical ultrasound image and video segmentation," Sensors, vol. 25, no. 8, 2025.

[12] M. Rainio et al., "Deep learning for medical ultrasound image segmentation," Journal of Digital Imaging, 2026.

[13] N. Alkhaldi et al., "Automating kidney disease diagnosis: A segmentation and classification approach using ultrasound imaging," Engineering, Technology & Applied Science Research, 2026.

[14] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," BMC Medical Imaging, vol. 15, no. 29, pp. 1–28, 2015.

[15] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," BMC Research Notes, vol. 15, no. 210, 2022.

[16] D. Müller, D. Hartmann, P. Meyer, F. Auer, I. Soto-Rey, and F. Kramer, "MISeval: A metric library for medical image segmentation evaluation," Studies in Health Technology and Informatics, vol. 294, pp. 389–390, 2022.

[17] S. Ostmeier et al., "USE-Evaluator: Performance metrics for medical image segmentation models with uncertain, small or empty reference annotations," arXiv preprint arXiv:2209.13008, 2022.

[18] Z. Zhang and U. Bagci, "Segmentation quality and volumetric accuracy in medical imaging," arXiv preprint arXiv:2404.17742, 2024.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. European Conference on Computer Vision (ECCV), Munich, Germany, 2018, pp. 801–818.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11531–11539.

[23] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in Proc. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11.

[25] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 574–584.

[26] S.-H. Chen, Y.-L. Wu, C.-Y. Pan, L.-Y. Lian, and Q.-C. Su, "Renal ultrasound image segmentation method based on channel attention and GL-UNet11," Journal of Radiation Research and Applied Sciences, vol. 16, p. 100631, 2023.

[27] M. I. Daoud, F. Abunameh, K. Shweikeh, S. K. Alzamer, M. Z. Ali, and R. Alazrai, "A comparative study of deep learning semantic segmentation models for kidney segmentation in ultrasound images using the Open Kidney Ultrasound Dataset," IEEE Access, vol. 13, pp. 144417–144433, 2025.

Learning Mathematics Through Play: Design and Development of a Serious Game for Undergraduate Students

Nur Nabilah Abdul Razak[1], Nur Nabila Farhana Mohd Noh[2], Nur Dayana Huda Mohd Zol Azhari[3], Ratna Zuarni Ramli[4]*, Ahmed M S Elaklouk[5], Azyan Yusra Kapi[6]
Universiti Teknologi MARA Negeri Sembilan Branch Seremban Campus, Seremban, Malaysia[1, 2, 3, 4]
School of Computing and Informatics, Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei Darussalam[5]
Universiti Teknologi MARA Johor Branch Pasir Gudang Campus, Johor, Malaysia[1, 6]

***Abstract*—Learning mathematics requires learners to have a good understanding of the concepts and problem-solving skills. However, mathematics is often associated with complexity and rarely perceived as connected to real-world applications. Educational games have the potential to make learning more enjoyable, motivating learners by offering challenges and control over their learning experience. Despite the growing research on mathematical games, existing educational games dominantly target primary and secondary levels, with limited focus on undergraduate calculus topics such as differentiation and integration. Visualizing mathematical concepts such as differentiation and integration through real-life scenarios and providing space for learners to learn while playing can be a suitable practice to persuade learners to learn mathematics. The aim of this study is to design and develop a game prototype that covers the topic of differentiation and integration. Method of instructional design: Analysis, Design, Development, Implementation, and Evaluation (ADDIE) is used to develop the game prototype, focusing on the systematic construction and expert validation. The serious game called CalcQuest Adventure was developed and distributed to eight experts to evaluate its suitability as a tool for learning the mathematical concepts of differentiation and integration. A verified instrument called the Suitability Evaluation Questionnaire (SEQ) is adapted in the study to evaluate CalcQuest Adventure. The result shows that the overall mean score is 3.85, which indicates positive feedback from experts. The findings contribute to the practical design insights for educators and developers by illustrating how real-life problems can be visualized and embedded into serious games for undergraduate calculus learning.**

***Keywords*—Mathematics education; game design; differentiation and integration; calculus; STEM**

I. Introduction

Mathematics helps to understand the complexities of the world, such as the patterns in nature, economic systems, and formulating predictions. Learning mathematics requires learners to have a good understanding of the concepts and problem-solving skills. Mathematical concepts explain how certain rules are applied and the reasoning behind them. Understanding the concepts enables learners to apply appropriate problem-solving strategies. This skill involves critical thinking that allows them to break down difficult problems into simpler parts. However, mathematics is often associated with complexity and disassociated to the real world [1, 2]. Thus, their perception is becoming one of the global issues, as many learners choose not to pursue mathematics. Convinced actions are required to change their perception and encourage learners to have more interest in learning mathematics. Creating an enjoyable learning environment has been shown to be one way to encourage learners to have a better understanding [3, 4]. Visualizing mathematical concepts in real-life scenarios and providing space for learners to learn while playing may support learners' engagement with mathematics. Educational games can make learning more enjoyable while motivating learners by offering challenges and control over their learning experience [3, 5]. Moreover, gaming is a popular activity among young people, and games appeal more to Millennials and Gen Z than movies [6]. Therefore, the use of game technology as an interactive learning media has the potential to eliminate boredom and fear of learning mathematics, especially for Millennials and Gen Z.

Even though educational games are an excellent teaching tool, there is a limited number of mathematics games that are designed for upper-level topics specifically for undergraduate students [4, 7]. Most mathematics games aim for pre-school to primary level that only cover basic calculation. Upper-level topics mostly presented in the form of textbooks, traditional video or slide presentation [8, 9]. Topics like differentiation and integration, which are often considered difficult topics are limitedly available in a more enjoyable form to learn [10]. Presenting the topic as the content of educational game and visualizing the topic to real-life scenarios may able to make learner comprehend easily. It has potential to help the learner to understand the mathematical concept of differentiation and integration and improve their problem-solving skills while playing. However, the studies that focus on designing and developing mathematics games for undergraduate students are limited [4, 7].

Despite the increasing adoption of educational games in mathematics education, most existing studies and applications continue to focus on basic mathematical skills at the primary and secondary levels. Research that addresses game-based learning for undergraduate calculus topics, particularly differentiation and integration, remains limited. This gap indicates a need for instructional tools that support conceptual understanding of

*Corresponding author

advanced mathematical topics through interactive and engaging learning environments.

Therefore, this study aims to explore the research activities in designing an educational game for undergraduate students, specifically on the topic of differentiation and integration. The primary objective of this study is to design and develop a serious game prototype for undergraduate calculus topics, namely differentiation and integration. It is represented by visualizing the mathematical concept interactively using graphics and animation by depicting a real-life scenario. The methodology used to develop the game prototype is the Analysis, Design, Development, Implementation, and Evaluation (ADDIE) method. This method is suitable for the research as the system requirements are clearly identified before development begins, reducing ambiguity and minimizing potential design errors. The iterative processes where continuous testing, validation, and improvement are significant for refinement. Evaluation of the suitability of the prototype as a learning tool using expert validation, contributing to the limited body of research on educational games for advanced mathematics topics. At this stage, this study focuses on design and suitability evaluation through expert validation and does not aim to measure student learning outcomes. At the end of this study, a discussion on the results of the study is explained, and some recommendations for future study are defined.

II. Previous Studies

Mathematics is a broad subject covering geography, science, politics, and society, and every aspect of life is affected by mathematics. It implies that mathematics is a critical need for sound work in life, irrespective of the field of study or profession of practice of anyone [2, 11]. Educational games, known as a combination of education and playful experience, can motivate the students to learn mathematics, which is mentioned to be a challenging subject [3, 12]. The games are increasingly used as educational resources for students because it is the most engaging and enjoyable virtual environment [10]. Despite the advantages of educational games, research emphasizes the scarcity of mathematics games specifically designed for advanced topics, particularly at the undergraduate level [10]. Most existing games target basic mathematical skills suitable for younger audiences, creating a gap in resources for more complex topics such as differentiation and integration, which are often perceived as challenging.

Motivation, self-confidence, and perceived usefulness of mathematics are some positive attitudes that correlate with learning outcomes. Previous studies highlight the importance of creating supportive learning environments that nurture positive attitudes [2]. Significantly learning by visualizing the equations into graphics and incorporating games into learning is one of the ways for providing a supportive learning environment [3, 13]. Mauntel & Zandieh [3] in their study found that combining game-based and dynamic geometry tools supports the transition of conceptual understanding to real-scenario application in learning mathematics. Their game, named Vector Unknown, created a motivating and exploratory environment, allowing students to experiment and strategize while learning vector concepts. The study shows that students developed richer and more flexible ways of thinking and learning mathematics.

Another perspective that is important in learning mathematics is looking at the view of the teacher. Observation from the teacher during the learning process can be a good input in creating a positive learning environment. According to Matic [7], who explored the perception of teachers on computer games, discovered that teachers generally view computer games as useful tools for enriching mathematics instruction. However, based on the analysis of their survey, most teachers clarify that there is a shortage of resources in terms of suitable games for learning. The teacher-centered research design revealed that future initiatives should focus on empowering teachers through training and developing constructivist game tools in teaching mathematics [7]. This finding is consistent with the research method and results shared by Salainti & Neman [13], where they found that computer games create a more enjoyable learning environment, with noticeable increase in student attention and enthusiasm during lessons. They also concluded that teachers are open and optimistic about integrating computer games in learning. Matic [7] and Salainti & Neman [13] ask teachers' opinions about the use of computer games in teaching subjects in classrooms and both studies applied teacher-centered research design approach.

Studies by Mauntel and Zandieh [3], Anagnostopoulou [10], Matic [7] and Salainti and Neman [13], have significant differences on methods used. Mauntel and Zandieh [3] and Anagnostopoulou [10], primarily focus on learners' engagement, motivation, and conceptual understanding through direct interaction with game-based environments. These studies applied student-centered design in evaluating the game using the Game Experience Questionnaire and Players' Experience of Need Satisfaction to capture students' responses during gameplay. While Matic [7] and Salainti and Neman [13] emphasize teachers' perceptions of usability, classroom feasibility, and instructional value rather than individual learner experience. Student-centered approaches provide rich insights into engagement and learning processes. However, this approach often fails to notice limitations related to curriculum alignment and classroom implementation. In contrast to teacher-centered approaches that are more focused on instructional integration and resource suitability, this study adopts the method, aiming to evaluate the suitability of a calculus-focused game prototype.

Calculus-focused games require learners to understand mathematical concepts through visualization or graphical representation that mimics a real scenario. A study that explores calculus-oriented games, such as the role-playing game developed by Anagnostopoulou [10], demonstrates a shift from procedural practice towards exploration and problem-based learning. Procedural practice in basic mathematics is frequently selected as the content of educational games designed for primary and secondary education levels. These games often emphasize repetitive practice, immediate feedback, and skill reinforcement, making them suitable for foundational learning, but less adaptable to the abstract reasoning required at higher levels of mathematics. Research on upper levels of mathematics games remains limited. This imbalance indicates that basic mathematics games are well established and calculus-focused games are still inadequate, emphasizing the need for further

studies that support higher-order mathematical thinking at the undergraduate level.

Not only are the resources of teaching mathematics using computer games limited, but the calculus topic, like differentiation and integration, has always been non-spotlight. Based on the search on the index-comprehensive and multidisciplinary academic research database, Web of Sciences in December 2023, there is a very limited amount of research done that is focused on teaching differentiation and integration topics. Anagnostopoulou [10] considers this gap and explores a role-playing game (RPGs) in mathematics education that includes the topic of differentiation and integration. The author aims to improve the students' engagement and understanding of the topic by designing an RPG and evaluating the game by adapting and devising Players' Experience of Need Satisfaction (PENS) and the Game Experience Questionnaire (GEQ). The result of the study shows a significant improvement in students' mathematical abilities and a positive attitude towards the subject, also responding to the gap of Matic [7] and Salainti & Neman [13]. The author also suggests that innovative approaches can be particularly effective in adult education settings, where traditional teaching methods may be less engaging.

The evaluation method applied by Anagnostopoulou [10] is similar to the method applied by Alsswey [14] and Rodrigues & Cheiran [15], but different in terms of questionnaire, as they used the Suitability Evaluation Questionnaire (SEQ) to evaluate their applications. SEQ consists of 14 questions that include the user's feelings while playing the game, and a fewer number of questions compared to GEQ and PENS. An optimized number of questionnaires may help respondents to be more focused on answers, committed to responding and reduce the burden on respondents. There are many arguments in the number of items in a questionnaire; however, the suitability of the questionnaire to the objective of the study is the most important factor to consider.

Comprehensive instruments such as the Game Experience Questionnaire and Players' Experience of Need Satisfaction are effective in evaluating the learners' immersive and motivational experiences, as applied by the research in [7] and [14]. However, these instruments are less suitable for early-stage prototype evaluation, where instructional suitability and usability are the main interests. SEQ has been applied in studies focusing on the practical assessment of educational applications, particularly in terms of ease of use, clarity of content, and overall appropriateness for instructional purposes [14, 15]. It is suitable to evaluate educational tools within limited time constraints. Therefore, SEQ is considered appropriate for this study, which adopts a teacher-centered design to evaluate the suitability of a serious game prototype for classroom integration rather than to measure learners' affective gameplay experiences.

Based on the input from previous studies, this research aims to fill in the gap by designing a prototype of a computer game that focuses on the calculus topic, which is differentiation and integration. To evaluate the design of game prototypes, a SEQ is employed through teacher-centered design research. The following section briefly describes the methodology.

III. METHODOLOGY

This study mainly focuses on the process of designing serious game content of real-life scenarios that relate to mathematical theories, which are the differentiation and integration topics. The aim of the game is to assist learners in understanding how differentiation and integration theories can solve real-life problems. The methodology of the study is divided into three phases, as described in Table I. The first phase is the Conceptual phase, which involves literature analysis on similar studies. The result of this phase is the theories and the content scope. The second phase is called the Design and Development phase, which involves specific steps that are frequently applied by application developers [16], which is Analysis, Design, Development, Implementation, and Evaluation (ADDIE). In this process, expert validation is needed to evaluate the low-fidelity design before the real game prototype is developed. A low-fidelity design is a paper sketch that has a game storyline that consists of real-life problems and can be solved by using appropriate mathematical theories. Three experts are appointed to examine the correctness of mathematics theories and how they apply to serious game design. The experts are appointed based on the criteria of having at least five years of experience in teaching calculus for undergraduate students. The third phase is the Game evaluation phase, which is carried out by developing a game prototype based on the finalized outcome from phase 2. Eight experts are appointed to evaluate the game prototype using SEQ to measure the suitability of the game to assist learners in learning differentiation and integration. The selection of experts for this phase followed the same criteria applied in the preceding phase. Accordingly, the three experts involved in the earlier evaluations were invited to participate in the assessment at this stage.

TABLE I. RESEARCH FRAMEWORK

Phase	Method	Expected result
Conceptual analysis	Literature study on Similar studies on existing mathematics games Mathematical concepts	Scope - differentiation and integration
Design and Development	ADDIE method Three experts to validate the low-fidelity design	Serious game prototype design
Game Evaluation	Eight experts to measure the suitability	Validated mathematics game prototype

A. Instrument

The instrument is adapted originally from Gil Gomez [17], called the Suitability Evaluation Questionnaire (SEQ). The questionnaire is used to evaluate the user experience with the product, such as satisfaction, acceptance, and security of use. The SEQ includes a short feedback questionnaire and focuses on the product's suitability for the users, consisting of 14 questions that include the user's feelings while playing the game. The questionnaire is scored from 1 to 5, with 1 indicating "not at all" and 5 indicating "very much", or 1 indicating "very easy" and 5 indicating "very difficult".

B. Expert Evaluation of Low Fidelity Design

Recruited experts are from the mathematics field who have good experience in teaching calculus, specifically differentiation and integration topics.

A low-fidelity design in the form of a paper sketch is given to the experts to evaluate the content of the game. Their responsibility is to ensure that all mathematical theories are correctly applied to specific real-life problems. They need to agree and disagree with 25 checklist items that describe the storyline of the game. Table II shows an example of questions. At the end of the checklist, experts can add further comments/ suggestions, and it is not compulsory. The size of the respondents is suitable for this study according to Paz [18].

TABLE II. EXAMPLE OF CHECKLIST ITEMS

Num	Question	Image Num	Theory	Agree (Yes/No)
1	A ladder slides down a tree. If the bottom moves away from the wall at 2 m/s, find how fast the top is falling when the bottom is 3 m from the wall and the ladder is 5 m long.	1a	Acceleration Differentiation	
2	If the ladder starts falling from rest and accelerates due to gravity, model the velocity function and find the acceleration at t=2 seconds.	1b	Acceleration Differentiation	
3	If the radius of the cylinder decreases over time, determine the time when the radius becomes half its initial value.	2a	Volume Differentiation	

C. Expert Evaluation of the Game Prototype

A total of eight mathematical experts are recruited for the study to evaluate the game prototype. The number of respondents is enough for this study, according to the heuristic evaluation method [18]. The SEQ is designed using Google Form and the survey link is shared to all experts together with a game installer. The respondent needs to play the game as long as they need, then they need to fill in the survey form. The result from the survey is analysed using descriptive analysis and a table of improvement actions is adapted from Ramli [19] and applied to categorize the results based on mean score. Table III shows the improvement action.

TABLE III. IMPROVEMENT ACTIONS BASED ON MEAN SCORES

Score	Action
0.00 - 1.66	Design must be improved
1.67 - 3.33	Design can be improved
3.33 - 5.00	Design is maintained

IV. RESULTS

This section describes briefly the result of the study on phase 2 and phase 3 that consists of: 1) suggested design of low-fidelity interface, 2) evaluation result for low-fidelity design and 3) evaluation result on game prototype using SEQ. Hence, the results are divided into two parts, namely design and evaluation of low-fidelity interface, and game evaluation result based on the research framework, as shown in Table I. The first phase of the methodology is carried out by utilizing the literature review process that concludes the scope of the research, which is differentiation and integration.

A. Design and Evaluation of Low-Fidelity Interface

A paper sketch of interfaces and a storyboard are designed using the ADDIE method. The paper sketches, also known as low-fidelity, are designed to mimic real-life scenarios purposely to present the real-world problem that can be solved by using mathematical theories, specifically differentiation and integration. Fig. 1 shows the example of low-fidelity interface designs, while Fig. 2 shows the design of interfaces for the homepage, menu, and game contents included in the example. For the interface, 1. Menu page and 2. Game map, shows the main pages that include the introduction of the game. For the interface, 3. Level 1 illustrates the real situation of a ladder leaning on a tree and to add a storyline to the game, there is a cat on top of the tree that needs to be saved. However, the ladder may slide from the tree, and this is where the differentiation theory, which is acceleration, is applied to describe how fast the ladder falls down. The players need to calculate using the differentiation (acceleration) theory by having certain conditions, like the height of the ladder and the tree. Interface Level 4 and Level 5 show the example of low-fidelity design for differentiation theory of volume and integration theory of heat.

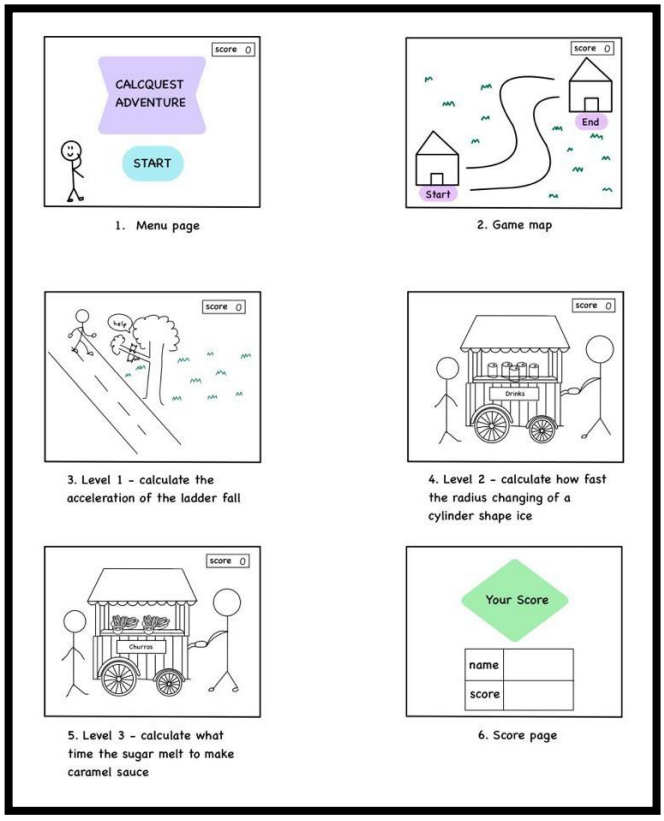


Fig. 1. Examples of low-fidelity interface designs for CalcQuest Adventure.

The recruited experts reviewed the description of each interface in the low-fidelity design and answered the questions by agreeing or disagreeing with the statements. The evaluation of low-fidelity design took around 3-6 weeks, and all experts

returned the feedback within acceptable time. Based on the feedback of all experts, it shows that there was no disagreement with any of the 25 statements in the checklist neither nor further arguments or suggestions from the experts. It can be concluded that all experts agree with the theory applied.

Based on the positive result of low-fidelity design evaluation, a game prototype is developed by using Scratch software and the duration of development took around 10 weeks. The game prototype is called CalcQuest Adventure and Fig. 2 shows the examples of game scenes. The main character, Jordyn, represents the player to experience the real-life scenario and handle all the challenges that need to be solved using the differentiation and integration theories. In order to make the game fun to play CalcQuest Adventure is designed by including game elements like game mechanics, goal, progression, balance, feedback, narrative and aesthetics. The different ways to collect points and answer the questions are examples of challenging fun factor that is included in the game design. For example, in Level 1, players engage in a shooting game for innovative question-solving. Level 2 introduces a catching game, which adds agility and coordination to the learning process. Lastly, Level 3 introduces a drag-and-drop game, requiring a hands-on approach to answering questions. This dynamic and engaging interface meets various learning styles, ensuring comprehensive educational experience at different skill levels.

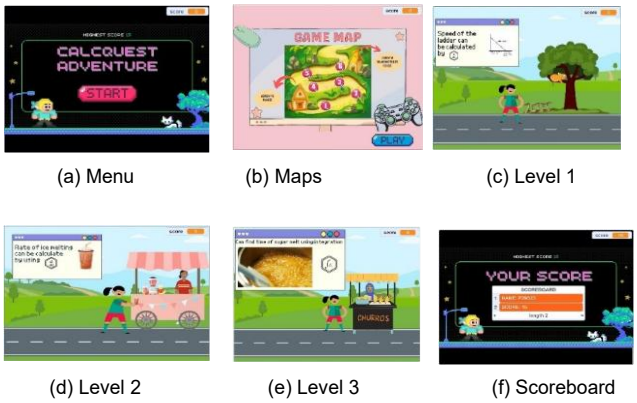


Fig. 2. Game scenes from CalcQuest adventure.

B. Game Evaluation Result

A total of eight mathematical experts participated in the study. The results from the survey are retrieved from the Google Form distributed within the game installer file. Table IV shows the results generated by SPSS 5.0. The overall mean score is 3.85 and referring to Table III, the design of CalcQuest Adventure falling under the “Design is maintained” category, where by the design is maintained and generally well-received by users with mean scores above 3.33 across most items. The highest value of mean score is 4.25 which is for item 8, addressing that the game is helpful for students’ learning. This result indicates that "CalcQuest Adventure" has the potential to become a valuable learning tool.

Items 7, 9, and 10 of the SEQ are negatively oriented, with higher scores reflecting increased discomfort or difficulty. Therefore, these items were reverse-scored during the analysis to ensure consistency in score interpretation across all instrument items, whereby higher scores uniformly represent more positive outcomes. Based on the analysis, the lowest value of the mean score is 3.00 for item 7, suggesting minor issues that may affect user comfort. Mean score 3.0 falls under category 2 which is the design that can be improved but not compulsory. No other item from the questionnaire has mean score less than 3.33 indicating CalcQuest Adventure giving positive user experiences. Aside from that, the total average of standard deviation from the eight mathematics experts is 0.7851 indicating a relatively low level of variability in the scores, the feedback consistent and stable performance across respondents.

Fig. 3 is a graphical representation of the relationship between mean values and the number of items, serving as a visual representation of the data derived from the SPSS analysis that is shown in Table IV. The visual graph provides a clearer visualization of the trends observed in the corresponding Table IV. It can be seen that almost all item distributed on the right side of the graph indicate the positive feedback from all experts.

TABLE IV. DESCRIPTIVE ANALYSIS

Item	Questions	Mean	Standard Deviation	Action
1	How much did you enjoy your experience with "CalcQuest Adventure"?	4.125	0.6409	Design is maintained
2	How immersed did you feel being in the environment of “CalcQuest Adventure"?	4.000	0.7559	Design is maintained
3	How successful were you in understanding through “CalcQuest Adventure"?	4.125	0.8345	Design is maintained
4	To what extent were you able to control “CalcQuest Adventure"?	3.875	0.8345	Design is maintained
5	How real is the virtual environment of “CalcQuest Adventure"?	3.750	0.8864	Design is maintained
6	Is the information provided by "CalcQuest Adventure" clear?	4.125	0.8345	Design is maintained
7	Did you experience any discomfort during your experience with “CalcQuest Adventure"?	3.000	0.9258	Design can be improved
8	Do you think that “CalcQuest Adventure" will be helpful for your students' learning?	4.250	0.7071	Design is maintained
9	Did you find the task difficult?	3.625	0.5175	Design is maintained
10	Did you find the “CalcQuest Adventure" interface difficult to use?	3.625	0.9161	Design is maintained

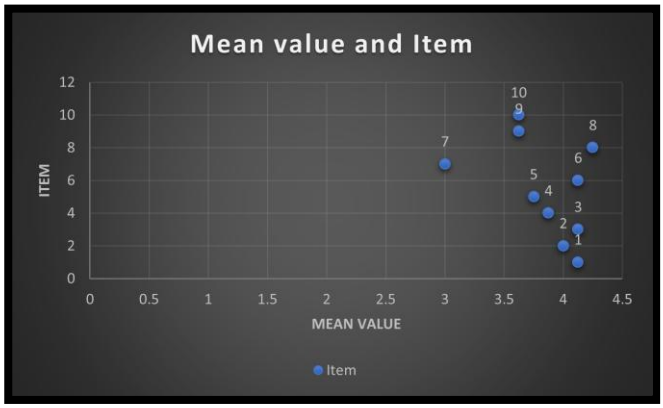


Fig. 3. Mean value vs. item

V. Discussion and Conclusion

This study aimed to design a mathematics game prototype for undergraduate students focused on enhancing the understanding of differentiation and integration. Mathematics is often perceived as abstract and disconnected from real-world, as it is frequently viewed by students as a challenging subject with limited practical relevance. This perception has contributed to a global decline in mathematics-engagement among undergraduate students. Thus, the use of game-based learning as a pedagogical tool to make mathematics more engaging, enjoyable, and applicable to real-world contexts can be the solution. However, limited resources of games designed for upper education level restrained the resolution. The goal of the study is to fill in the gap by providing a game that covers the topic of calculus specifically differentiation and integration and relate the mathematics theories to real-life problems.

CalcQuest Adventure, a game prototype developed using the ADDIE instructional design model is designed to mimic the real-life problem and players solve the problems using the theories of differentiation and integration. The game prototype underwent an expert evaluation phase twice, first during the early phase by creating a low fidelity interface design in the form of paper sketch and second during the last phase, where eight experts are recruited to assess the design using a validated instrument known as the Suitability Evaluation Questionnaire (SEQ).

The feedback data, presented in the form of mean scores and standard deviations, offers valuable insights into the perceived strengths and limitations of the game. The evaluation comprised ten items, each rated on a 5-point Likert scale. Nine out of ten evaluation items received mean scores above 3.34, confirming that the game's design is well-maintained across several dimensions. Experts reported high enjoyment (item 1), a strong sense of immersion (item 2), and clear instructional content (item 3). These results suggest that CalcQuest Adventure succeeds in making mathematics more relatable and engaging, a critical step in addressing student disengagement from the subject. Furthermore, the game was perceived as helpful for students' learning (item 8), highlighting its potential as a supplementary educational tool.

One notable exception was the item regarding user discomfort, which scored a mean of 3.00. Although this value still falls within an acceptable range, it indicates that some aspects of the user experience may require further refinement to improve player comfort. It is very crucial to choose evaluation method and a significant type of evaluation that cater to usability must be carried out in order to get the detail issues on this matter [20]. However, this score does not undermine the overall merit of the game. The standard deviation values, ranging from 0.51 to 0.93, reflect moderate variability among expert opinions, suggesting general consensus while acknowledging some differing perspectives [21]. These findings are important for iterative improvements to the game design.

In summary, the expert evaluation supports the conclusion that CalcQuest Adventure provide valuable insights into the quality and effectiveness of a game design, particularly during the prototype and formative evaluation stages. By leveraging experts' knowledge and experience, the SEQ helps identify design strengths and weaknesses before conducting large-scale user testing specifically on the topics of differentiation and integration. Based on the result of analysis, the game prototype offers a learning experience that may enhances students' understanding on the topic. With an overall mean score of 3.85, the results indicate expert approval, validating the game's readiness for classroom implementation. Previous study on education field shows that range of 3.41 to 4.20 reflects a high level of engagement and satisfaction [18]. Higher levels of engagement and satisfaction in learning may contribute to improved learning outcomes and academic performance [22].

By improving comfort and minimizing usability issues, may enhance the game's effectiveness. According to experts, CalcQuest Adventure demonstrates that game-based learning can serve as a tool in learning mathematics for upper education level. Further research needs to be done to improve the design of the game and more suitable evaluation has to be carried out in order to create a suitable tool to learn mathematics, especially for undergraduate students. Future work may explore both content refinement and game engine optimization, particularly through the integration of more advanced algorithms to improve overall system performance and user experience [23]. This study contributes to the limited body of research on serious games for undergraduate calculus by demonstrating the feasibility of expert-validated game design for differentiation and integration

Declaration

All data analyzed during this study are included in this research. The authors declare that there is no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

The corresponding author conceived the research idea and provided overall guidance. All authors contributed equally to the research activities, game design and development and data analysis. Manuscript writing was led by the corresponding author, with discussion and brainstorming involving all co-authors. All authors read and approved the final manuscript.

Acknowledgment

We thank the Universiti Teknologi MARA for the academic support of this research.

REFERENCES

[1] Manaf, A. (2021). Student attitude and mathematics learning success: A metaanalysis. International Journal of Instruction, 14(4), 209–222. https://doi.org/10.29333/iji.2021.14413a

[2] Krath, J., Schürmann, L., & von Korflesch, H. F. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games, and game-based learning. Computers in Human Behavior, 125, 106963. https://doi.org/10.1016/j.chb.2021.106963

[3] Mauntel, M., Zandieh, M. Forms of Structuring Space by Linear Algebra Students with Video Games and GeoGebra. Int. J. Res. Undergrad. Math. Ed. 10, 754–780 (2024). https://doi.org/10.1007/s40753-024-00246-2

[4] Li, Y., Chen, D., & Deng, X. (2024). The impact of digital educational games on student's motivation for learning: The mediating effect of learning engagement and the moderating effect of the digital environment. PloS ONE, 19(1), e0294350. https://doi.org/10.1371/journal.pone.0294350

[5] The Star. (2023). Young People Spend More Time Playing Video Games Than Watching Movies. The Star Online. https://www.thestar.com.my/tech/tech-news/2023/07/22/young-people-spend-more-time-playing-video-games-than-watching-movies

[6] Rosni, N. H., Anuar, N., & Idris, A. S. (2024). The effectiveness of educational games in learning mathematics among secondary school students. Malaysian Journal of Computing, 9(2), 1–12.

[7] Matic, L. J., Karavakou, M., & Grizioti, M. (2023). Is digital game-based learning possible in mathematics classrooms?: A study of teachers' beliefs. International Journal of Game-Based Learning (IJGBL), 13(1), 1-18.

[8] Lisnani, Putri, R. I. I., Zulkardi, & Somakim. (2023). Web-based realistic mathematics learning environment for 21st-century skills in primary school students. Journal on Mathematics Education, 14(2), 253–274. https://doi.org/10.22342/jme.v14i2.253-274

[9] Dele Ajayi, O., Lawani, A. O., & Adeoye, F. A. (2019). The critical role of mathematics in everyday life and its implications for teaching. International Journal of Mathematics Education, 14(3), 45–58.

[10] Anagnostopoulou, E. (2024). Digital Role-Playing Games as a Novel Approach to Learning Mathematics: A Quantitative Analysis. Adults Learning Mathematics, 18(1), 49-69.

[11] Shirkarami, A., Abadi, K. A., Ardakani, S. P., & Azimi, S. (2020). The effect of mobile games on math learning of third graders of elementary schools. In Proceedings of the 2nd International Serious Games Symposium, ISGS 2020 (pp. 126-130). Article 9375439 (Proceedings of the 2nd International Seriou11

[12] Shirkarami, A., Abadi, K. A., Ardakani, S. P., & Azimi, S. (2020). The effect of mobile games on math learning of third graders of elementary schools. In Proceedings of the 2nd International Serious Games Symposium, ISGS 2020 (pp. 126-130). Article 9375439 (Proceedings of the 2nd International Serious Games Symposium, ISGS 2020). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ISGS51981.2020.

[13] Salainti, E., & Neman, M. I. E. (2024). Teachers' Perception of Using Digital Game-Based Learning in Teaching and Learning Process. EDUKASIA: Jurnal Pendidikan Dan Pembelajaran, 5(1), 955-962. https://doi.org/10.62775/edukasia.v5i1.904

[14] Alsswey, A., Malak, M. Z., & El-Qirem, F. A. (2025). Effect of virtual reality on perceptions of usability, suitability, satisfaction, and self-efficacy among architecture and design university students. Architectural Science Review, 68(1), 56-64.

[15] Rodrigues, A., & Cheiran, J. F. P. (2020). Virtual look around: suitability evaluation for virtual tour in multiple platforms. In Simpósio de Realidade Virtual e Aumentada (SVR) (pp. 337-346). SBC.

[16] Kemouss, H., Abdennour, O., Erradi, M., & Khaldi, M. (2023). The ADDIE pedagogical engineering model: From analysis to evaluation. In Handbook of research on scripting, media coverage, and implementation of e-learning training in LMS platforms (pp. 42-70). IGI Global.

[17] Gil-Gómez, J., Bigné, E., & Moliner, M. A. (2013). Suitability Evaluation Questionnaire (SEQ): A tool for assessing user satisfaction in educational games. Computers & Education, 67, 146–156. https://doi.org/10.1016/j.compedu.2013.03.013

[18] Paz, F., Paz, F. A., Pow-Sang, J. A., & Collazos, C. (2018). A formal protocol to conduct usability heuristic evaluations in the context of the software development process. Int. J. Eng. Technol, 7(2.28), 10-19.

[19] Ramli, R. Z., Sahari Ashaari, N., Mat Noor, S. F., & Ismail, M. (2024). Designing a mobile learning application model by integrating augmented reality and game elements to improve student learning experience. Education and Information Technologies, 29, 1981–2008. https://doi.org/10.1007/s10639-023-11874-7

[20] Shear, B. R., & Reardon, S. F. (2021). Using pooled heteroskedastic ordered probit models to improve small-sample estimates of latent test score distributions. Journal of Educational and Behavioral Statistics, 46(1), 3-33.

[21] Yi, F. S., & Kutty, F. M. (2023). A Study of Primary School Students in Johor: The Role of Temperament and Teacher-Student Relationships in Student Engagement. International Journal of Academic Research in Progressive Education and Development, 12(2).

[22] Ramli, R. Z., Ramli, A. H., Osman, N., Ismail, Z., Musa, M., & Satari, H. (2021). Online learning: Insight into Malaysia University lecturers' perspective. International Journal of Academic Research in Business and Social Sciences, 11(2), 1250-1258.

[23] Kapi, A. Y., Sunar, M. S., & Abd Algfoor, Z. (2020, December). Summary of Pathfinding in Off-Road Environment. In 2020 6th International Conference on Interactive Digital Media (ICIDM) (pp. 1-4). IEEE.

Applying the AuRa Consensus Model for Digital Certificate Management in a Private Ethereum Blockchain

Robiah Arifin[1*], Wan Aezwani Wan Abu Bakar[2*], Mustafa Man[3*],
Mohamad Afendee Mohamed[4], Evizal Abdul Kadir[5]
Department of Big Data, Infostructure and Network Management Centre,
Universiti Sultan Zainal Abidin, 21030 Kuala Nerus, Terengganu, Malaysia[1]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, 22200 Besut Campus, Besut, Terengganu, Malaysia[2, 4]
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia[3]
Faculty of Engineering, Universitas Islam Riau, Jl. Kaharuddin Nasution 113, Pekanbaru 28284, Riau, Indonesia[5]

***Abstract*—The issue of fake certificates has been widely identified, and their prevalence has increased significantly in recent years. This growing trend has become a global concern due to its adverse impact on educational standards. A key factor contributing to the problem is the continued reliance on manual processes for issuing and verifying certificates. To address these challenges, this study proposes the use of an authority round (AuRa) consensus algorithm for managing certificate data on the Ethereum blockchain. AuRa, a member of the proof of authority (PoA) family, facilitates consensus among nodes distributed across multiple servers and networks. This mechanism plays a vital role in preserving the integrity and decentralization of the blockchain while ensuring the security of transactional data. Furthermore, the study investigates how AuRa enables efficient certificate data transactions within a private Ethereum environment. It also evaluates the algorithm's performance in terms of transaction speed per second (TPS) and throughput per second (TGS), demonstrating its effectiveness for managing certificate transactions on a blockchain network. Then the TPS and TGS results substantiate the suitability of AuRa for digital certificate generation, evidenced by its stable and efficient performance within a controlled private server environment.**

***Keywords*—Blockchain technology; fake certificate prevention; Authority round (AuRa) algorithm; ethereum private network; Proof of Authority (PoA); certificate verification system**

I. Introduction

Certificates are important for graduate students as proof that they have completed their studies. Unfortunately, numerous job seekers use fake certificates to apply for jobs. Fake certificates were first traced in the 19th century [1]. Such certificates were issued by a nonexistent university.

Fake certificates have become a global issue and have been detected worldwide [2]. In response to the scenario, the research proposes the digital certificate using blockchain technology supported by the authority round (AuRa) consensus algorithm.

Blockchain allows participants to write and read the transactions in the network [3]. To ensure the security of blockchain, this research implements the Ethereum and AuRa consensus algorithms. The consensus algorithm serves as a security mechanism by establishing agreement among nodes to ensure the throughput is consistent, not different, and valid. This research also implemented the smart contract as a logic flow to transact data into the blockchain.

The research aims to create a digital certificate using the AuRa consensus algorithm through a smart contract and to measure the transaction speed per second (TPS) and throughput per second (TGS). The implementation is executed on a private server that consists of three datasets, each containing a different number of records and including data information such as student number, gender, and faculty. This research aims to address the issue of counterfeit certificates by generating a digital certificate using the Ethereum blockchain supported by the AuRa consensus algorithm.

The contribution of this study lies in demonstrating that the AuRa consensus mechanism is capable of supporting blockchain-based transaction processes. The effectiveness of these processes is evaluated using throughput performance indicators, namely TPS and TGS, in the context of digital certificate generation. Furthermore, the findings indicate that AuRa can effectively address issues related to the prevention of counterfeit certificates in the digital certificate issuance process.

Section II outlines the issue of counterfeit certificates and examines the underlying causes. It also provides an overview of blockchain technology, with a focus on the Ethereum platform and the AuRa consensus algorithm, highlighting their potential in addressing the problem of certificate forgery.

Section III details the proposed implementation of the method to mitigate these issues. Section IV presents the results and analysis of the implementation, followed by Section V, which discusses the findings and acknowledges the limitations of the study. Finally, Section VI concludes the study by summarizing the implementation of blockchain technology and the AuRa consensus algorithm for certificate generation, and outlines potential directions for future research.

A. Novelty

This study introduces a novel real-world deployment of the AuRa consensus algorithm for digital certificate generation, offering an empirical assessment that is largely absent in prior

*Corresponding author.

blockchain-based certificate management literature. While existing works commonly focus on conceptual frameworks or simulated environments, this research advances the field by evaluating AuRa within an operational private server setting and by utilizing two concrete performance metrics, TPS and TGS. The use of TPS and TGS as paired performance indicators provides a measurable and systematic way to assess both transactional efficiency and data-handling capability, thereby contributing a practical and reproducible evaluation methodology to the research domain.

B. Contributions

Then the findings demonstrate that AuRa delivers stable and favorable TPS and TGS performance outcomes, substantiating its suitability for real-world digital certificate generation. This work contributes empirical evidence that AuRa can reliably support certificate issuance workflows with predictable throughput and consistent operational behaviour within a private server environment.

Beyond confirming performance viability, the study provides a validated deployment model that institutions can adopt or extend for secure certificate management. Collectively, these contributions highlight the practical applicability of AuRa and position this study as one of the few that bridges the gap between theoretical blockchain designs and deployable certificate management solutions.

II. Literature Review

The main criterion for employers to choose an employee is the certificate. Employers verify certificates manually, which is time-consuming and labor-intensive. The difficulty of verifying the certificate contributes to the issue of the production of fake certificates. The problem of forged certificates persists among the higher education community [4]. Fake certificates have become a global issue, which is associated with certificate scams.

Numerous factors cause these problems. One of the factors is an inefficient student management system. The highly trustworthy system needs to be implemented to eliminate the fake certificate problem [5]. Another factor is that employers often do not complete the certificate verification process because it requires extra human resource efforts [6]. Manual verification processes also exacerbate these problems [7]. Employers must contact academic institutions manually for certificate verification, which takes significant time and effort [8].

To address the growing issue of fake certificates, blockchain technology can provide a solution. Blockchain technology is an effective method to address these problems. Many applications and methods based on a blockchain framework have been proposed to prevent fake certificate issues, as described in Table I.

Based on Table I, show the implementation of blockchain for a digital certificate. However, these studies face performance issues for the verification and authentication of digital certificates. Unfortunately, this study also did not discuss the measurement and performance results of the digital certificate.

TABLE I. Digital Certificate Management Using Blockchain

Authors	Case Study
Fang et al. (2024)[9]	Credential verification
Feng et al. (2024)[10]	Certificate authentication
Liu et al. (2024)[11]	Certificate verification
Mai et al. (2024)[12]	Certificate verification
Samarai (2024)[13]	Certificate verification

A. Blockchain Technology

Blockchain is a decentralized, distributed database used for storing transaction data and information. Blockchain was introduced as a new method of managing and storing data, replacing traditional centralized systems with decentralized systems [14]. The data storage consists of a chain of records known as blocks.

Data is stored across different nodes on different servers. The blockchain is also known as a peer-to-peer (P2P) network [15] because the nodes communicate with each other through a P2P network. Blockchain uses a storage system to store digital data and protect it with cryptography and a hash function. Cryptographic enforcement in blockchain networks ensures immutability, rendering the recorded logistics data resilient against tampering and hacking [16].

In 1991, the capabilities of blockchain were improved with the secure chain of blocks, which could be implemented through the cryptographically secure method described by Haber and Stornetta [17]. Today, public blockchains are widely available through various readily available platforms to be implemented by researchers. However, the popular and widely used ones are Bitcoin, Ethereum, and Hyperledger [18]. Ethereum was established years back, focusing on encouraging businesses to adopt the blockchain in their business activities [19]. To date, these blockchain platforms share similarities but also retain their unique features.

Both Bitcoin and Ethereum are public blockchains that allow everybody to participate in writing and reading the transactions in the network. This concept is known as the permissionless blockchain. Participants are allowed to connect and disconnect from the network at any time without permission.

B. Ethereum

Ethereum was first introduced by Vitalik Buterin in 2013. It was intended to overcome the limitations of Bitcoin. In 2014, Ethereum's development was crowdfunded in 2014 to improve the technology. Ethereum has implemented distributed data storage that allows everyone to run their own applications on the blockchain [20]. It is capable of storing data with flexible block sizes. The Ethereum engine requires support from Ethereum blockchain clients.

The Ethereum client can be deployed on development and production environments, but it has difficulties and needs expertise to set up and configure the environment. Parity and Geth are the popular and active supporters of the blockchain community [21].

However, the Ethereum clients need to be supported by a consensus algorithm to ensure the security and performance of transactions in the blockchain.

C. Consensus Algorithm

A consensus algorithm is an agreement process among nodes across different servers and networks. It supports and commits a blockchain system to maintain the consistency and decentralization of nodes [22]. Consensus algorithms enhance the security of blockchain transactions by ensuring that throughput remains consistent and valid. The consensus algorithm is functional not only in terms of security aspects, but it also provides services for reducing the execution time of transaction processes and providing the highest throughput.

The performance of consensus depends on its algorithm and protocol for executing transactions in the blockchain. Its performance also directly impacts the processing speed and scalability of the blockchain. A consensus algorithm that commits transactions in less time is more efficient. Faster consensus execution increases throughput. In addition, various consensus algorithms support blockchain transactions. The most popular and widely implemented blockchain applications are proof of authority (PoA), practical byzantine fault tolerance (PBFT), proof of stake (PoS), and proof of work (PoW). Table II shows the consensus and its specialty.

TABLE II. CONSENSUS AND SPECIALTY

Consensus	Specialty
PoS	More spikes in blockchain
PoW	Able to solve a mathematical puzzle
Proof of Luck	Implemented the random selection
Proof of Elapsed Time	Able to set up the scheduling while the timeout.
Proof of Space	Implemented in bigger size of hard disk environment
PoA	Well-suited for public blockchains deployed in private environments.

D. PoA

PoA was introduced by Wood as an alternative to PoW and PoS. PoA was introduced to overcome the limitations of PoW and PoS [23]. Therefore, the PoA has become a popular consensus algorithm because of its high-performance transaction process and security. It is able to process thousands of transactions per second without impacting its security. PoA also provides fault tolerance and efficiency.

Besides that, the PoA consensus also needs less energy and fewer computational resources to implement. It also makes efficient use of network bandwidth. This algorithm is used in permissioned blockchains and is implemented in two versions, namely AuRa and Clique. These versions contain their own advantages and disadvantages.

This research implemented the AuRa because it was originally proposed for a private network environment [24]. That is why this research chose to implement AuRa to overcome the fake certificate issue. Additionally, nearly 4,000 blockchain projects across various domains, including education, management, healthcare, and insurance, have utilized AuRa [25]. As proven, AuRa is more secure compared to Clique, and AuRa throughput is higher than PoW.

Although this study does not formally benchmark alternative consensus algorithms or certificate verification systems, the selection of the AuRa is justified based on its compatibility with permissioned blockchain environments. Unlike PoW, which is energy-intensive and better suited for public networks, or PoS, which introduces complexity in validator selection, AuRa offers deterministic block production, low energy consumption, and predictable performance. These characteristics align well with the operational needs of academic certificate issuance, where trust is centralized, and efficiency is prioritized. Future work may involve empirical benchmarking against other consensus mechanisms to further validate this selection.

III. RESEARCH METHODOLOGY

To overcome the fake certificate issues, this study proposed to implement the AuRa consensus algorithm to create the digital certificate. Fig. 1 shows the framework of implementation of AuRa to create the digital certificate.

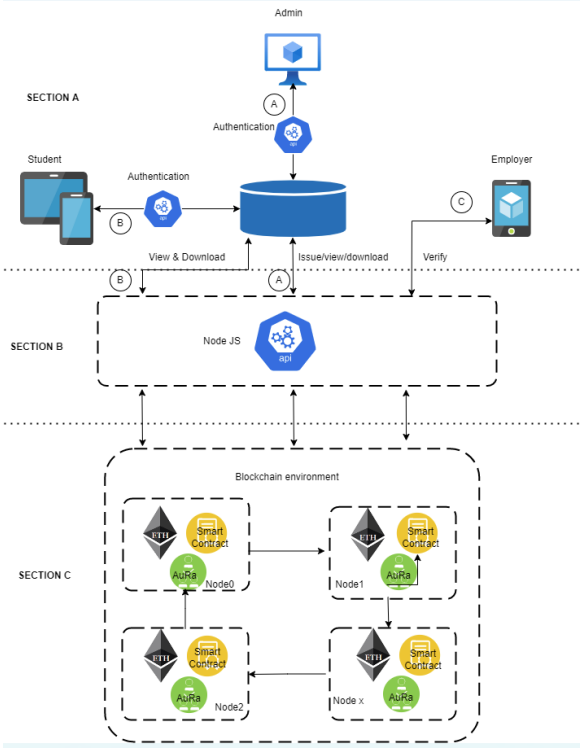


Fig. 1. Implementation of the AuRa framework.

Regarding Fig. 1, the framework to implement the AuRa consensus algorithm is divided into three sections. Section A focuses on the application layer, which comprises three primary roles: administrator, student, and employer. The administrator is responsible for issuing digital certificates through the system, ensuring that each record is correctly generated, authenticated, and securely committed to the blockchain. This role also incorporates access control mechanisms to prevent unauthorized issuance and ensures that certificate data is transmitted and stored in a tamper-resistant manner.

Students, as legitimate certificate holders, are permitted to securely access, view, and download their digital certificates via authenticated sessions within the application. Security controls implemented at this layer, such as identity verification and encrypted communication, ensure that only the rightful student can retrieve their certificate information. Employers function as external verifiers who authenticate the validity of certificates through the application's verification module. This module leverages blockchain-backed immutability to ensure that verification results cannot be falsified, providing a trusted mechanism that prevents manipulation, forgery, or unauthorized alterations of certificate records. Collectively, these security functions establish a secure, transparent, and tamper-resistant application environment.

Section B refers to the backend processes, which serve as the core operational layer supporting the application layer. The backend utilizes an application programming interface (API) to transmit and receive commands and data between the application layer and the blockchain network. These APIs manage the execution of certificate issuance requests, retrieval of certificate data, and verification queries while ensuring secure, authenticated, and structured communication between the frontend and the underlying blockchain infrastructure.

To enhance security, the backend incorporates multiple protective mechanisms. All data transmissions are encrypted to prevent interception or tampering during communication. Input validation and authentication checks are enforced for every request to mitigate unauthorized access and potential injection attacks. Transaction records are logged in an immutable audit trail to provide non-repudiation and support forensic analysis if discrepancies occur. Furthermore, the backend enforces role-based access control, ensuring that only authorized administrators, students, or employers can execute specific operations, thereby maintaining the integrity, confidentiality, and availability of the digital certificate system. These combined security measures ensure that the backend can reliably manage certificate issuance, retrieval, and verification while minimizing potential threats.

Lastly, Section C, the blockchain layer forms the foundational infrastructure of the digital certificate system, providing tamper-resistant storage, consensus validation, and secure transaction recording. This layer leverages the AuRa consensus algorithm to ensure that all certificate issuance and verification transactions are validated by a network of authorized nodes, maintaining integrity and consistency across the distributed ledger. The blockchain layer interacts with the backend via secure APIs, receiving transaction requests for certificate issuance, verification queries, and data retrieval, and returning confirmation or result status to the application layer.

Security functions at the blockchain layer are critical to maintaining trust and reliability. Each transaction is cryptographically signed and recorded in an immutable block, ensuring non-repudiation and protection against data alteration. Consensus mechanisms prevent single-node manipulation, while periodic block validation guarantees that any conflicting or unauthorized transactions are rejected. Access control is enforced at the node level, restricting operations to authorized validators.

Additionally, the blockchain maintains redundancy across nodes, enhancing system availability and resilience against node failures or malicious attacks. Combined with the secure backend and application layers, the blockchain layer ensures a comprehensive, end-to-end security framework for the digital certificate management system, enabling reliable issuance, verification, and auditability of certificates.

Lastly, the framework is comprehensively limited to certificate generation and verification activities. To guarantee the authenticity, security, and reliability of the produced certificates, both activities are required to systematically progress through the processes outlined in Sections A, B, and C.

A. AuRa Consensus Algorithm

The AuRa consensus algorithm was proposed for the Parity client. Parity is one of Ethereum's clients that applied Rust as a programming language. It has been implemented by Lava, VecHain Thor, xDai DPOS network, Microsoft Azure (for deployment only), and Kovan Testnet. This consensus became popular because it was widely adopted in blockchain applications.

This consensus assumes the authorities are honest and the network is synchronous among the nodes. Besides that, AuRa consists of four steps, including transaction pending, proposed block, voting, and as depicted in Fig. 2.

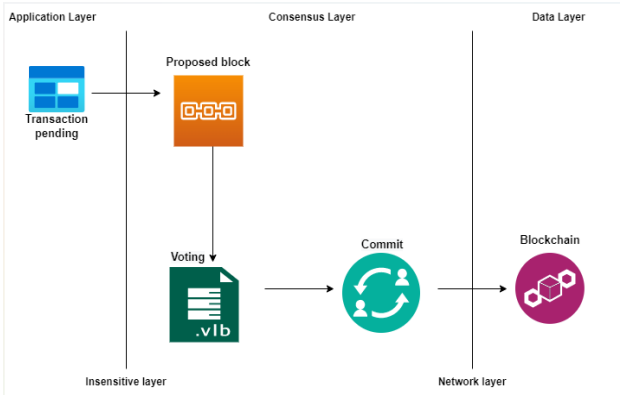


Fig. 2. The steps of the AuRA process.

In Fig. 2, the AuRa algorithm starts by assigning a leader; each authority has its turn to assign as a leader using the Round Robin Algorithm. Then the leader will propose a block, and the authorities will decide if the proposed block is accepted or not. The decision or voting process is based on Eq. (1).

$$2f + 1 \leq n \quad (1)$$

Regarding Eq. (1), where f is a faulty node, n is the number of nodes. In that case, the number of faulty nodes must be at least as many as the number of nodes 50%. Additionally, only trusted authorities are able to add the block to the blockchain. Trusted authorities that call n number and implement the majority among the authorities able to create a node.

After the leader is assigned, the leader will propose the block in the block proposal. Then the voting process will execute on block acceptance, an authority's members will vote on the block proposed by a leader, whether that block is able to commit or

not. If the leader or node is not valid, the process assign new leader will reprocess.

B. Create and Configure AuRa Engine

The AuRa engine comprises the Ethereum blockchain. Once the AuRa engine was installed, the Ethereum blockchain was automatically installed. Each AuRa engine was installed on a different server and is called a node. This research created the nodes at private servers. Each node will act as an authority on the network and issue blocks in the transaction process. Fig. 3 shows the creation of the AuRa engine. AuRa engine created at port xxxx:xxxx, located in the parity folder. These engines are connected with the xxxx.toml file.

Coding 1: Create the AuRa engine

```
[1]  run –rm -ti  -p xxxx:xxxx
[2]  -v ~/.local/share/parity/docker/:/home/parity/
[3]  .local/share/parity/ parity/parity:v3.0.0
[4]  --config /home/parity/.local/share/parity/xxxx.toml
[5]  --jsonrpc-interface all
```

Fig. 3. Create AuRa engine.

Then, after completing the coding to create the AuRa engine, the next step is to create the authority account, as in Fig. 4. The authority account created at port xxxx:xxxx is similar to the port at the AuRa engine. The location and toml file are also similar to the AuRa engine.

Coding 2: Create authority account

```
[1]  run –rm -ti  -p xxxx:xxxx -v
[2]  ~/.local/share/parity/docker/:/home/parity/
[3]  .local/share/parity/ parity/parity:v3.0.0
[4]  --config/home/parity/.local/share/parity/xxxx.toml
```

Fig. 4. Create an authority account.

Based on the execution of Fig. 4, the results in Fig. 5 illustrate the authority account successfully created. The authority account that was created is 0x008ce3c5316c23c8fdee747b0ae59188879xxxxx.

Result : Create authority account

```
[1]  {"jsonrpc":"2.0","result":"0x008ce3c5316c23c8fdee747b0ae59188879xxxxx","id":0}
```

Fig. 5. Result achieved.

C. Smart Contract

Smart contracts were proposed by Nick Szabo in the early 1990s. A smart contract is a platform and a high-level programming language to execute the mining process automatically [26]. The few available programming languages for smart contracts are Solidity, Mutan, Viper, and Chaincode.

The function of smart contracts is to support the Ethereum blockchain. It was developed and stored in the Ethereum blockchain. A smart contract is a programmable code that executes predefined commands stored on a blockchain. These commands include attributes, conditions, and rules that must be met before data can be recorded on the blockchain. Smart contracts leverage programmed logic to autonomously enforce access control lists and system constraints defined within the blockchain infrastructure [27]. In this research, a smart contract was implemented to facilitate the execution of commands for storing data on the blockchain.

Additionally, this research implemented Solidity as a programming language to develop the smart contract code. The smart contract was developed and tested on the Remix integrated development environment (IDE) platform. After the smart contract was successfully tested, it was rewritten on the Truffle framework. At the truffle framework, the smart contract was running using the truffle command to migrate the smart contract and the AuRa consensus algorithm.

Before executing the migration process, the configuration between nodes and the truffle framework must be considered. Fig. 6 captures the coding of the configuration of the nodes and the truffle framework. Line 1 indicates the server to create the AuRa engine. It consists of information about the host (line 3), port (line 4), and network ID at line 5.

Configuration nodes and truffle framework

```
[1]  Networks:{
[2]  Development: {
[3]  Host:"xxx.xx.x.x"
[4]  Port: xxxx,
[5]  Network_id:"xxxxx"
[6]  },
[7]  }
```

Fig. 6. Configuration of nodes and truffle.

Additionally, after configuring the nodes and the truffles framework as depicted in Fig. 6, the next step is to integrate both nodes and truffles. Used the truffle init command, the nodes and truffles will migrate as illustrates at Fig. 7.

Fig. 7. The migration nodes and truffles.

The smart contract address is crucial for integrating transaction data into the blockchain. Additionally, the migration process generates the application binary interface (ABI) file, which is essential for the transaction process.

Smart contracts play a pivotal role in the blockchain ecosystem by enabling decentralized automation that eliminates the reliance on traditional intermediaries. Through this cryptographic execution, these protocols secure data integrity and ensure absolute immutability against unauthorized

tampering. Furthermore, the embedded algorithmic logic allows for the precise enforcement of operational conditions and strict access control mechanisms across the network.

This automated governance enhances system transparency and auditability, providing an immutable ledger of verifiable transaction histories. Consequently, smart contracts serve as the foundational backend infrastructure that drives the functionality of decentralized applications and modern digital economies. Any modifications made to the smart contract logic do not affect system operations, provided that the updated contract undergoes the requisite node reconfiguration and truffle deployment processes.

D. Integration and Implementation

Before certificate issues, Python was implemented as a programming language to generate the certificate that embeds the AuRa consensus algorithm and smart contract. To ensure the AuRa consensus algorithm and smart contract are able to communicate with each other, the Web3 library is implemented. Web3, commonly known as web3.py, is a Python library that is able to interact with smart contracts and Blockchain.

It's commonly implemented in decentralized applications (dApps) to assist the sending transactions process, communication with smart contracts, writing and reading the block data in JavaScript Object Notation (JSON) format, to incorporate the solution for fake certificate issues. The certificate information was converted into JSON format to transact data into the blockchain through the AuRa consensus algorithm. The important data, such as student number, name, result, program, and faculty, were stored in a blockchain database. These data are important and must be accurate and reliable to avoid the fake certificate issues.

To evaluate the framework's practical applicability, the experimental setup simulates a multi-institution infrastructure by deploying a private Ethereum network across several servers. These nodes are strategically distributed across two geographically distinct locations to measure crucial network metrics, including reliability, synchronization overhead, and transaction speeds under realistic network latencies. This distributed private network closely mirrors a real-world, cross-organizational deployment where multiple independent authorities participate in the certificate verification process. Within this environment, digital certificates were generated and evaluated using three distinct datasets, each encompassing varying data types to test the system's robustness against realistic workloads.

Additionally, to ensure integrity and immutability, a private key is utilized in the generation of a digital certificate. A private key is a randomly generated set of numbers used in cryptographic processes. In this research, private keys are generated for selected users who possess the authority to create digital certificates. This research also grants administrators the explicit authority to revoke private keys upon theft or loss. To further prevent unauthorized administrative exploitation, key deployment is restricted to a single-user concurrency model.

E. Hardware and Software

To implement the proposed method in this research, several servers are required. Specifically, four servers were utilized to establish individual nodes, with each server physically located at a distinct on-premises location. The implementation relies on the AuRa consensus algorithm and the Ethereum blockchain engine, both of which were installed and configured on every node across all servers.

Additionally, a smart contract was developed, which was installed and configured on only one of the nodes. Other supporting software tools, including Truffle, Remix IDE, and Ganache, were accessed via their online versions, eliminating the need for local installation on the servers.

IV. Result and Analysis

This study aims to demonstrate that the AuRa algorithm is executable and effective for resolving certificate management. Within the context of digital certificate management, TPS and TGS serve as sufficient and primary evaluation metrics; TPS directly indicates the system's throughput capacity during high-volume issuance periods, while TGS reflects the responsiveness and latency of real-time certificate verification. For a rigorous performance, AuRa was implemented on the myCert platform using authentic graduation certificate data to generate and validate its operational suitability under realistic workloads.

MyCert is an in-house application that manages digital certificates, including the generation and verification process. It uses Python as its programming language and PyCharm as its Integrated Development Environment (IDE). The AuRa consensus algorithm is installed and embedded on myCert platform.

A. Geographic Location

The study deployed identical datasets across three geographically dispersed locations connected through a wide area network (WAN) topology, with each site hosting a data instance linked through intermediary network nodes. This distributed configuration reflects real-world environments in which data must be transmitted and validated across multiple infrastructures. The datasets were processed during peak operational periods to evaluate system behaviour under high-load conditions, specifically to measure the performance of the AuRa consensus mechanism in terms of TPS and TGS. This approach enables a more precise assessment of how network congestion and distributed processing influence overall performance within a WAN-based architecture.

TPS quantifies the number of successful transactions processed per second. It aims to measure the performance of AuRa, focused on transaction speed per second. This study used three sets of data: set 1 consists of 1,021 data, set 2 includes 2,435 data, and set 3 comprises 3,422 data. Three sets of data refer to the number of graduation student starting 2023 to 2025.

Based on three sets of data, the AuRA is used to generate the certificate. The performance of AuRa is measured by TPS and TGS. Fig. 8 illustrates the result of TPS. Based on the TPS result at different locations, the Data Set 1 is 31.59, Data Set 2 is 29.69, and Data Set 3 is 18.59. The same location for TPS result is 40.32,32.03 and 19.25.

Regarding Fig. 8, it is observed that the number of transactions significantly impacts the system's throughput performance, measured in TPS. However, the TPS results across

these datasets are not consistent, indicating that factors beyond transaction volume, such as network latency or processing overhead, may influence overall performance. This research also measured the TGS. The TGS represents the total size of data that transactions process successfully per second. This research measures the TGS for the size of the data.

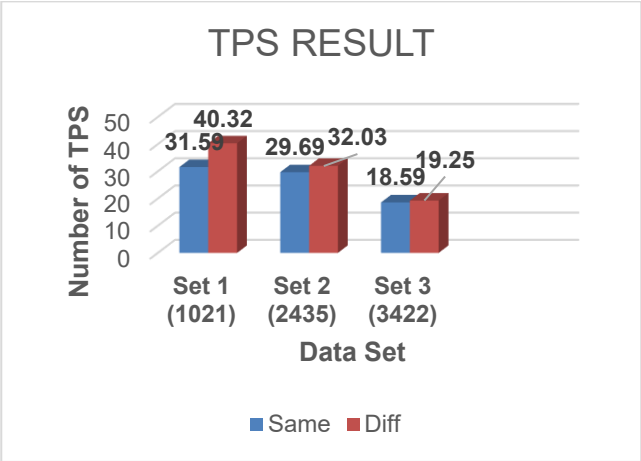


Fig. 8. The TPS result.

Regarding data in Table II, the TGS results were measured and compared with the execution at different and the same locations. Fig. 9 states the result of a number of TGS. Based on the TGS results, it showcases that set 1 has successfully transacted 1,064,266.14 KB per second, while set 2 is 992,814.83 KB per second, and lastly, set 3 successfully transacted 624,364.33 KB per second for a different location. The same location for TGS result is 1,358,194.53 KB per second, 1071194.95 KB per second, and 655,649.11 KB per second.

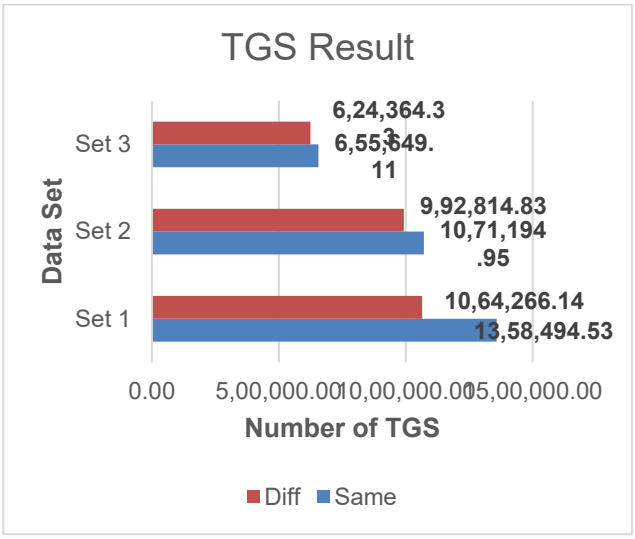


Fig. 9. Number of throughputs per second.

B. Number of Nodes

The number of nodes can significantly impact the performance of a consensus algorithm. In this study, the AuRa consensus algorithm was executed to generate digital certificates under three different scenarios using the same dataset.

In the first scenario, three nodes were deployed; in the second, four nodes; and in the third scenario, five nodes were implemented. To ensure a controlled testing environment, all three configurations were evaluated using uniform network bandwidth, identical server capacities, and a standardized dataset volume. This setup was designed to evaluate the effect of node scalability on system performance.

The TPS results presented in Fig. 10 show that the first scenario, which deployed three nodes, achieved a throughput of 53.32 transactions per second, outperforming the second scenario with four nodes, which recorded a TPS of 50.92. In the third scenario, where five nodes were implemented, the TPS increased to 52.25, indicating a partial recovery of performance relative to the four-node setup. These observations demonstrate that node count exhibits a non-linear influence on system throughput, suggesting that merely increasing the number of authority nodes does not guarantee improved performance within the AuRa consensus environment.

However, this experiment is limited to evaluating only three configurations, including three, four, and five nodes, under a three-dataset, which restricts the depth of performance interpretation. The analysis does not yet account for several critical factors that could further affect TPS behaviour.

Specifically, the study does not evaluate system performance when an authority node fails or behaves maliciously, which is essential for understanding AuRa's robustness in adversarial conditions. Similarly, the experiment does not explore the impact of increased block time or varying network latency, both of which can significantly influence throughput in distributed blockchain deployments.

Furthermore, the scalability of the system beyond five nodes, especially in multi-institution or multi-university settings, remains unexamined. Addressing these dimensions in future work would provide a more comprehensive assessment of AuRa's performance characteristics and operational resilience in large-scale, heterogeneous environments.

The TPS results presented in Fig. 10 indicate that the performance in the first scenario, which utilized three nodes, achieved a throughput of 53.32 transactions per second. This is higher compared to the second scenario, which involved four nodes and recorded a TPS of 50.92. Interestingly, in the third scenario, where five nodes were implemented, the TPS increased to 52.25, demonstrating improved performance relative to the second scenario. These findings suggest that the number of nodes has a non-linear effect on system performance, where increasing the number of nodes does not necessarily result in higher throughput.

Based on the analysis presented in Fig. 10, it can be concluded that the number of nodes has a measurable impact on the performance of the AuRa consensus algorithm. Inconsistencies in the empirical results are largely driven by synchronization delays inherent in multi-node deployments. A critical factor is the temporal overhead required for all nodes to achieve state synchronization, which inherently influences overall throughput and latency.

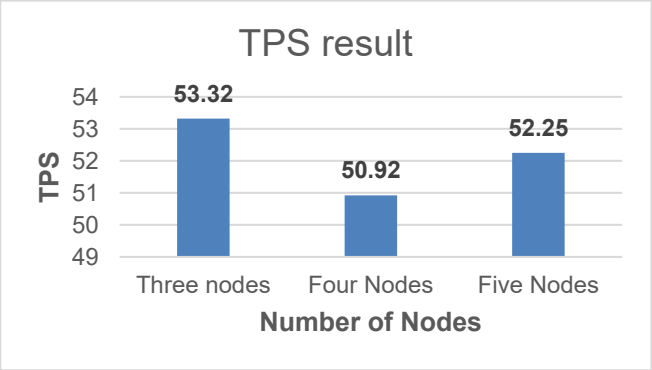


Fig. 10. Comparison of TPS results based on the number of nodes.

C. Security Analysis

In the context of security, low TPS performance in the AuRa consensus mechanism introduces several potential security vulnerabilities that merit careful consideration. When each transaction requires extended processing time, the prolonged block finalisation window increases exposure to adversarial interference. Specifically, a malicious leader may exploit this delay to insert falsified data, manipulate timestamps, or execute equivocation attacks before finality is reached.

Such conditions weaken the integrity guarantees of the protocol and elevate the probability of inaccurate or unauthorized certificate records being committed to the blockchain. In extreme cases, these delays may disrupt the synchronous assumptions underlying AuRa, thereby increasing the risk of block production inconsistencies.

Similarly, TGS performance indicates inefficiencies in handling data throughput, which may further exacerbate security risks within the certificate generation workflow. When the system struggles to transfer or commit data at a stable rate, node desynchronization becomes more likely, particularly in networks with multiple validators. Such desynchronization can create opportunities for adversaries to exploit temporary inconsistencies between nodes, potentially resulting in chain forks or interrupted block propagation.

If a malicious node initiates an attack during these vulnerable intervals, such as injecting altered certificate information, the resulting chain divergence could compromise system reliability and undermine trust in the blockchain-based certificate infrastructure. Therefore, maintaining adequate TPS and TGS performance is not merely an efficiency requirement but a critical component of ensuring the security and correctness of the AuRa-driven system

V. Discussion and Limitation

This study implemented AuRa to create digital certificates. The implementation focused on the measurement of TPS and TGS in only two situations: the same location and different locations of nodes. However, the study did not record the average gas usage per transaction during certificate creation.

A. Consensus Algorithm

This research implements only one consensus algorithm, which causes a limitation in evaluating and measuring the TPS and TGS results. If more than one consensus algorithm had been implemented in this study, comparisons could provide a stronger and more comprehensive evaluation.

B. Server

This study implements the AuRa algorithm to generate certificates on a private server. Thus, at least three servers are needed to establish the nodes. The private server implementation might incur extra costs and resources due to distinct configurations and maintenance requirements.

VI. Conclusion and Future Work

This work contributes a real-world deployment of a blockchain-enabled digital certificate management system built on a private Ethereum network using the AuRa consensus algorithm. Moving beyond conventional blockchain implementations, this study offers empirical evidence of its practical viability in strengthening certificate integrity and resolving verification-related challenges. Crucially, the core technical contribution lies in optimizing block generation and transaction throughput within a private network architecture, thereby effectively mitigating certificate forgery while addressing the performance and scalability bottlenecks inherent in standard consensus mechanisms.

Blockchain allows read and write access to the data. To ensure the security of blockchain, it is supported by the AuRa consensus algorithm. The AuRa consensus algorithm is able to create a secure environment within the transaction process until the transaction is committed to the blockchain. This research also implemented the AuRa consensus algorithm, Ethereum, and smart contracts, which have been embedded with Python coding.

The implementation of the certificate data is converted into JSON format. Then, execute the experiment, including three sets of data on a private server. However, the evaluation reveals inconsistent results on TPS and TGS. Regarding these results, geolocation, size of data, and number of nodes affected the TPS and TGS results.

Moreover, enhanced performance in TPS and TGS strengthens the system's capability to detect and prevent the use of fake or invalid certificates. By enabling real-time verification and rapid detection of discrepancies within the blockchain-based certificate infrastructure, the system reduces the risk of certificate forgery or duplication. This efficiency allows for continuous monitoring and immediate rejection of suspicious or tampered credentials, thereby upholding data integrity and trust in the digital certification ecosystem. Ultimately, the improved performance ensures not only operational effectiveness but also reinforces the security and reliability of the entire certificate management process. Regrettably, a direct comparative analysis with prior studies within the same domain could not be established. This limitation arises because existing literature fails to explicitly detail its empirical results and the specific performance metrics employed.

For future work, this study attempts to generate certificates under various conditions, including different numbers of nodes. The results are then compared and analyzed to determine whether the number of nodes has a significantly affects TPS and TGS.

ACKNOWLEDGMENT

We express sincere gratitude to the Center of Research and Innovation Management (CREIM) at UniSZA for their invaluable financial support towards the publication and the Research Group of Intelligent Data (ID). Our heartfelt appreciation also goes to all team members from UniSZA, including Pn. Robiah Arifin, the PhD candidate for the technical configuration of the project, and Dr. Wan Aezwani Wan Abu Bakar, and Assoc. Prof. Dr. Mohamad Afendee as her Supervisor and Co-Supervisor for the conceptual and technical proofreading, as well as the UMT member, Assoc Prof. Ts. Dr. Mustafa Man, for the network linkages and the reviewer's suggestions. Also, we are grateful for the international contributions and collaboration provided by Dr. Evizal Abdul Kadir, a research fellow from Universitas Islam Riau, Indonesia.

REFERENCES

[1] Carmichael, J. J., & Eaton, S. E. (2023). Fake degrees and fraudulent credentials in higher education: Conclusions and future directions. In Fake Degrees and Fraudulent Credentials in Higher Education (pp. 269-285). Cham: Springer International Publishing.

[2] Arifin, R., Wan Abu Bakar, W. A., Man, M., & Kadir, E. A. (2024). Performance Evaluation of the AuRa Consensus Algorithm for Digital Certificate Processes on the Ethereum Blockchain. International Journal of Advanced Computer Science & Applications, 15(11).

[3] Huynh-The, T., Gadekallu, T. R., Wang, W., Yenduri, G., Ranaweera, P., Pham, Q. V., ... & Liyanage, M. (2023). Blockchain for the metaverse: A review. Future Generation Computer Systems, 143, 401-419.

[4] Thakare, N., Narad, S., & Surysvanshi, Y. (2024, December). Fake certificate detection by using blockchain. In AIP Conference Proceedings (Vol. 3188, No. 1). AIP Publishing.

[5] Aini, Q., Harahap, E. P., Santoso, N. P. L., Sari, S. N., & Sunarya, P. A. (2023). Blockchain based certificate verification system management. APTISI Transactions on Management, 7(3), 191-200.

[6] Tumati, T. V., Tian, Y., & Jiang, X. (2024, January). A soulbound token certificate verification system (sbtcert): Design and implementation. In 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0345-0350). IEEE.

[7] Pu, Shuyi, and Jasmine Siu Lee Lam. "The benefits of blockchain for digital certificates: A multiple case study analysis." Technology in Society 72 (2023).

[8] Rustemi, A., Dalipi, F., Atanasovski, V., & Risteski, A. (2023). A systematic literature review on blockchain-based systems for academic certificate verification. IEEE Access.

[9] Fang, J., Feng, T., Guo, X., Ma, R., & Lu, Y. (2024). Blockchain-cloud privacy-enhanced distributed industrial data trading based on verifiable credentials. Journal of Cloud Computing, 13(1), 30.

[10] Feng, X., Wang, L., Bai, X., & Yang, P. (2024). Distributed identity management mechanism based on improved block-chain certificateless encryption algorithm. Physical Communication, 65, 102341.

[11] Liu, H., Ming, Y., Wang, C., Zhao, Y., Zhang, S., & Lu, R. (2024). Blockchain-assisted verifiable certificate-based searchable encryption against untrusted cloud server for Industrial Internet of Things. Future Generation Computer Systems, 153, 97-112.

[12] Mai, C. K., Iqbal, M. S., Rohith, A., Suchetan, T. C. K., & Shinde, P. C. (2024). Applicant Credentials Tracker for Employment Using Blockchain Technology. International Journal of Intelligent Systems and Applications in Engineering, 12(3s), 320-327.

[13] Samarai B.A. (2024).Use of Blockchain Technology in Educational Field.International Journal on Technical and Physical Problems of Engineering,15(4),140-151

[14] Rani, P., Sachan, R. K., & Kukreja, S. (2024). Educert-chain: a secure and notarized educational certificate authentication and verification system using permissioned blockchain. Cluster Computing, 27(7), 10169-10196.

[15] Howell, A., Saber, T., & Bendechache, M. (2023). Measuring node decentralisation in blockchain peer to peer networks. Blockchain: Research and Applications, 4(1), 100109.

[16] Villegas-Ch, W., Gutierrez, R., Govea, J., & García-Ortiz, J. (2025). Integrated AI, IoT, and blockchain for enhancing security and traceability in perishable logistics. Emerg. Sci. J, 9, 2471-2496.

[17] Tripathi, G., Ahad, M. A., & Casalino, G. (2023). A comprehensive review of blockchain technology: Underlying principles and historical background with future challenges. Decision Analytics Journal, 9, 100344.

[18] Pancari, S., Rashid, A., Zheng, J., Patel, S., Wang, Y., & Fu, J. (2023). A systematic comparison between the ethereum and hyperledger fabric blockchain platforms for attribute-based access control in smart home IoT environments. Sensors, 23(16), 7046.

[19] Mai, C. K., Iqbal, M. S., Rohith, A., Suchetan, T. C. K., & Shinde, P. C. (2024). Applicant Credentials Tracker for Employment Using Blockchain Technology. International Journal of Intelligent Systems and Applications in Engineering, 12(3s), 320-327.

[20] Arifin, R. B., Bakar, W. A. W. A., Man, M. B., & Kumar, B. P. (2024). Tackling counterfeit certificate problems with blockchain technology: a review. International Journal of Advanced Technology and Engineering Exploration, 11(119), 1392.

[21] Arifin, R., Bakar, W. A. W. A., Man, M., & Kadir, E. A. (2025). Comparative Performance Analysis of Original AuRa and Improved AuRa Consensus Algorithms in Chain Hammer Digital Certificate Simulation. International Journal of Advanced Computer Science & Applications, 16(10).

[22] Islam, M. M., & In, H. P. (2023). Decentralized Global Copyright System Based on Consortium Blockchain with Proof of Authority. IEEE Access. https://doi.org/10.1109/SCC55611.2022.00054

[23] Hu, Y., Tian, G., Jiang, A., Liu, S., Wei, J., Wang, J., & Tan, S. (2023). A practical heartbeat-based defense scheme against cloning attacks in PoA blockchain. Computer Standards & Interfaces, 83, 103656.

[24] Sarfaraz, A., Chakrabortty, R. K., & Essam, D. L. (2023). The implications of blockchain-coordinated information sharing within a supply chain: A simulation study. Blockchain: Research and Applications, 4(1), 100110.

[25] Zhang, Xinrui, Rujia Li, Qin Wang, Qi Wang, and Sisi Duan. "Time-manipulation attack: Breaking fairness against proof of authority Aura." In Proceedings of the ACM Web Conference 2023, pp. 2076-2086. 2023. https://doi.org/10.1145/3543507.3583252

[26] Wu, G., Wang, H., Lai, X., Wang, M., He, D., & Chan, S. (2024). A comprehensive survey of smart contract security: State of the art and research directions. Journal of Network and Computer Applications, 226, 103882.

[27] Zaw, T. O. K., Anbananthen, K. S. M., Muthaiyah, M. S., Balasubramaniam, B., Kannan, R., Kalid, K. S., ... & Mohammad, S. (2025). A Systematic TOGAF-Driven Framework for Blockchain-Based Food Traceability with Access Control Lists. HighTech and Innovation Journal, 6(2), 461-475.

A Bibliometric Analysis of Internet of Things and Augmented Reality Applications in Agriculture: A Scopus-Based Review (2019–2026)

Ruziana Mohamad Rasli, Sobihatun Nur Abdul Salam
School of Multimedia Technology and Communication, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

***Abstract*—The integration of Internet of Things (IoT) and Augmented Reality (AR) technologies has emerged as a transformative approach in modern agriculture, enabling precision farming, real-time monitoring, and immersive decision-support systems. This study presents a bibliometric analysis of research trends related to IoT and AR applications in agriculture using data retrieved from the Scopus database. A total of 72 publications were gathered from 2020 to 2026. The metadata is then analyzed using bibliometric techniques to identify publication trends, document types, research themes, and emerging directions. The results reveal a steady increase in publications, with the highest number of records (20 documents) in 2024 and a slight drop to a total of 12 documents in 2025. Conference papers represent the largest proportion of publications (20 documents), followed by book chapters (17 documents), and reviews with 13 documents, reflecting the changing and experimental nature of this research area. Key research themes include precision agriculture, smart farming systems, digital twins, sensor networks, and real-time visualization for crop monitoring. The findings highlight the increasing role of AR interfaces integrated with IoT sensor data for enhanced agricultural decision-making. This study provides an inclusive overview of research developments and identifies potential future directions, including AI-driven AR visualization, digital agriculture ecosystems, and sustainable farming technologies.**

***Keywords*—Bibliometric analysis; Internet of Things; Augmented Reality; smart agriculture; precision farming; Scopus database**

I. Introduction

Agriculture is undergoing rapid transformation with the adoption of digital technologies designed to improve productivity, sustainability, and resource efficiency. Among these technologies, the Internet of Things (IoT) and Augmented Reality (AR) have attracted significant attention in recent years. IoT enables real-time data collection through networks of sensors, while AR provides interactive visualization that allows users to interpret complex data within a real-world environment.

IoT-based agricultural systems typically utilize sensors to monitor parameters such as soil moisture, temperature, humidity, nutrient levels, and crop health. These systems enable farmers to make data-driven decisions and automate agricultural processes. However, the visualization and interpretation of large volumes of sensor data remain challenging for farmers and agricultural practitioners.

Augmented Reality has emerged as a complementary technology capable of overlaying digital information onto the physical environment. When combined with IoT, AR can provide immersive visualization of farm data, enabling users to monitor crop conditions, analyze sensor readings, and manage agricultural systems more efficiently. The integration of AR and IoT in agriculture supports several applications, including:

1) Precision farming: Precision farming refers to a data-driven agricultural approach that uses IoT sensors, GPS, and analytics to monitor field variability and optimize farming practices. It enables farmers to apply water, fertilizers, and pesticides precisely where needed, improving efficiency and crop yield.

Recent studies show that IoT-enabled sensors provide real-time data on soil conditions, plant health, and environmental factors, allowing targeted interventions such as optimized irrigation and fertilization [1], [2]. This significantly reduces resource wastage and enhances productivity.

2) Crop monitoring: Crop monitoring involves continuous observation of crop conditions using IoT sensors, drones, and imaging technologies. It helps detect issues such as plant stress, diseases, and nutrient deficiencies at early stages.

IoT-based systems collect real-time environmental and crop data, enabling farmers to track crop growth and respond quickly to changes [1]. Advanced systems can monitor parameters like soil moisture, temperature, and pest activity, improving overall farm management.

3) Agricultural training and education: AR plays a major role in agricultural training by providing interactive and immersive learning environments. It allows farmers and students to visualize farming processes, machinery operations, and crop conditions without physical risks.

Recent studies highlight that AR-based training systems improve knowledge transfer, enhance understanding, and support skill development in agriculture by simulating real-world scenarios [3]. This is especially useful for training in modern farming technologies.

4) Smart irrigation management: Smart irrigation uses IoT sensors and automated systems to optimize water usage based on real-time environmental data. It ensures that crops receive the right amount of water at the right time.

Research shows that IoT-based irrigation systems can significantly reduce water consumption while maintaining optimal soil moisture levels through automated control mechanisms [2]. These systems are particularly important in addressing water scarcity and climate variability.

5) Digital twin farming environments: Digital twin technology creates a virtual replica of a physical farm using real-time data from IoT sensors. This allows farmers to simulate, monitor, and optimize agricultural processes. Recent studies indicate that digital twins integrate IoT, AI, and AR technologies to model crop growth, predict outcomes, and support decision-making [4], [5]. They enable advanced applications such as yield prediction, disease detection, and scenario simulation.

IoT and AR technologies support several applications in agriculture, including precision farming, crop monitoring, agricultural training, smart irrigation, and digital twin environments. Precision farming enables data-driven decision-making through real-time sensor data [1], [2], while crop monitoring improves early detection of environmental changes [1]. AR enhances agricultural training through immersive visualization [3], and smart irrigation optimizes water usage using automated systems [2]. Furthermore, digital twin technology enables simulation and predictive analysis for improved farm management [4], [5].

Despite the growing interest in this interdisciplinary field, comprehensive reviews focusing specifically on the convergence of IoT and AR in agriculture remain limited. Bibliometric analysis provides a systematic method for evaluating scientific publications, identifying research trends, and mapping the development of a research field.

The objectives of this study are:

- To analyze publication trends related to IoT and AR in agriculture.
- To identify the dominant document types and publication sources.
- To explore emerging research themes in this field.
- To identify potential research directions for future studies.

Therefore, this study aims to analyze global research trends related to IoT and AR in agriculture using bibliometric techniques based on publications indexed in the Scopus database.

The next section will focus on the literature on the digital transformation in agriculture, the usage of Internet of Things (IoT) in agriculture, the usage of Augmented Reality (AR) in agriculture, and the integration of both technologies in agriculture, followed by methodology, results, discussion, research challenges and future directions, and conclusion section.

II. Literature Review

A. Digital Transformation in Agriculture

The agricultural sector is undergoing a significant transformation driven by digital technologies aimed at improving productivity, sustainability, and resource efficiency. Traditional farming methods, which rely heavily on manual labor and experience-based decision-making, often result in inefficient resource use and limited adaptability to environmental changes [6]. These limitations have accelerated the transition toward smart agriculture, where technologies such as the Internet of Things (IoT), Augmented Reality (AR), and artificial intelligence (AI) are increasingly adopted.

Recent studies highlight that smart agriculture systems integrate sensing, communication, and data analytics technologies to enable real-time monitoring and automation [7]. Among these technologies, IoT and AR play complementary roles in data acquisition and visualization, forming the foundation of modern precision agriculture systems.

B. Internet of Things (IoT) in Agriculture

The Internet of Things (IoT) has become a core component of smart agriculture by enabling real-time monitoring through interconnected sensors and devices. IoT systems collect environmental data such as soil moisture, temperature, and humidity, allowing farmers to make data-driven decisions and optimize resource utilization [6].

1) Usage of IoT in agriculture: IoT applications in agriculture cover a wide range of smart farming activities that help improve productivity and sustainability. For example, smart irrigation systems use sensor data to optimize water consumption by supplying water only when needed, reducing waste and improving efficiency. In addition, soil and environmental monitoring systems utilize wireless sensor networks to collect real-time data on factors such as soil moisture, temperature, humidity, and nutrient levels. IoT technology is also applied in livestock tracking and health monitoring, allowing farmers to monitor animal location, behavior, and health conditions more effectively. Furthermore, automated greenhouse management systems help control environmental conditions such as lighting, temperature, and ventilation to support optimal crop growth. IoT can also assist in crop disease detection and yield prediction through continuous data collection and analysis. Overall, these applications enable farmers to make more accurate data-driven decisions, leading to increased agricultural productivity, reduced resource wastage, and more sustainable farming practices.

2) Advantages of IoT: IoT offers several important benefits in agriculture by enabling smarter and more efficient farming practices. One of the main advantages is real-time monitoring and control of farming conditions, allowing farmers to continuously observe environmental factors such as temperature, humidity, soil moisture, and nutrient levels. This capability helps improve resource efficiency by reducing unnecessary use of water, fertilizers, and energy, ultimately minimizing wastage and operational costs. In addition, IoT

technologies contribute to enhanced crop yield and quality through accurate data collection and timely interventions. IoT also supports the automation of repetitive agricultural tasks, such as irrigation, fertilization, and greenhouse management, reducing labor dependency and increasing productivity. Furthermore, these technologies play a significant role in supporting precision agriculture techniques, where farming activities are optimized based on real-time data and specific crop requirements.

3) Disadvantages of IoT: Despite its many advantages, the implementation of IoT in agriculture also faces several challenges. One of the major limitations is the high initial setup and maintenance costs, which can be difficult for small-scale farmers to afford. In addition, IoT systems heavily depend on reliable internet connectivity, particularly for real-time monitoring and data transmission, making deployment difficult in rural areas with limited network coverage. Data security and privacy concerns also arise due to the continuous collection and storage of sensitive agricultural data. Furthermore, the technical complexity of IoT systems requires users to possess certain technical skills and knowledge for effective operation and maintenance. Another challenge involves the integration of IoT technologies with existing farming systems and traditional agricultural practices, which may require additional infrastructure and adaptation efforts.

C. Augmented Reality (AR) in Agriculture

AR is emerging as a powerful tool for enhancing data interpretation in agriculture by overlaying digital information onto the physical environment. Recent research positions AR as an important interface that connects IoT, AI, and other smart farming technologies [8].

1) Usage of AR in agriculture: AR applications in agriculture provide interactive and immersive solutions that improve farm management, monitoring, and training processes. One important application is the real-time visualization of crop and soil data, where farmers can view information such as moisture levels, temperature, and crop conditions directly through AR interfaces. AR is also used in farm management systems through mobile devices or wearable technologies, enabling users to access agricultural data and control systems more efficiently in the field. In addition, AR supports agricultural training and simulation environments by creating realistic virtual experiences that help farmers and students learn farming techniques and equipment operations safely and effectively. Another application involves equipment maintenance guidance, where AR overlays provide step-by-step visual instructions for repairing and maintaining agricultural machinery. Furthermore, AR technology can be used for digital twin visualization of agricultural systems, allowing users to observe and interact with virtual representations of farms and farming processes in real-time.

AR allows users to interact with data directly in the field, improving situational awareness and decision-making.

2) Advantages of AR: AR offers several important benefits in agriculture by improving the way farmers interact with and interpret agricultural data. One major advantage is its ability to enhance understanding through visual and interactive data representation, allowing users to view farming information more clearly and intuitively. AR also improves efficiency in farm operations by providing real-time guidance and visual support during monitoring, maintenance, and management activities. In addition, AR facilitates training and knowledge transfer by creating immersive learning environments that help farmers and agricultural workers better understand farming techniques and equipment usage. Another important benefit is the reduction of cognitive load when interpreting complex datasets, as AR simplifies data visualization through interactive overlays and graphical representations. Furthermore, AR supports real-time decision-making by enabling farmers to access accurate information instantly, leading to faster and more effective agricultural management.

3) Disadvantages of AR: Despite its many advantages, the implementation of AR in agriculture also encounters several challenges. One of the primary limitations is the high development and deployment costs associated with AR applications and hardware, which may not be affordable for many farmers, especially small-scale agricultural operators. In addition, the limited availability of AR devices and supporting infrastructure in rural areas can restrict wider adoption. AR systems also face usability challenges in outdoor agricultural environments, where factors such as sunlight, weather conditions, and field terrain may affect visibility and system performance. Furthermore, AR applications heavily depend on device performance and battery life, which can limit continuous operation during farming activities. Another challenge is the need for user training and adaptation, as farmers and agricultural workers may require time and technical knowledge to effectively use AR technologies in daily agricultural practices.

D. Integration of IoT and AR in Agriculture

The integration of IoT and AR represents a major advancement in smart agriculture by combining real-time data collection with immersive visualization. IoT sensors continuously gather environmental data, while AR interfaces transform this data into interactive visual formats that enhance decision-making. Recent studies show that integrating IoT with AR and digital twin technologies improves monitoring, visualization, and farm management efficiency [1]. Furthermore, AR-based systems integrated with IoT sensors significantly enhance user interaction and support more effective agricultural decision-making [9].

These technologies support various advanced agricultural applications that improve farming efficiency and decision-making. One important application is precision farming with real-time visual analytics, where sensor data is combined with AR visualization to help farmers monitor and manage crops more accurately. This integration also supports smart irrigation monitoring and control by providing real-time information on water usage, soil moisture, and irrigation performance. In

addition, IoT and AR technologies support crop health and disease visualization, allowing farmers to identify plant conditions and potential diseases more effectively through interactive visual displays. Another significant application is the development of digital twin farming environments, where virtual representations of farms are used for monitoring, simulation, and analysis. Furthermore, the integration of these technologies improves decision support systems by presenting real-time data and visual insights that support farmers in making faster and more informed agricultural decisions. By bridging the gap between data acquisition and interpretation, IoT-AR systems improve operational efficiency and enable proactive agricultural management.

E. Comparison with Traditional Farming

Traditional farming relies on manual observation and experience-based decisions, which often lead to inefficiencies and delayed responses to environmental changes [6]. The transition from traditional farming to IoT and AR-based agriculture highlights significant improvements in efficiency, accuracy, and sustainability. IoT and AR-based systems enable real-time monitoring, automation, and data-driven decision-making for users (see Table I).

TABLE I. COMPARISON WITH TRADITIONAL FARMING

Aspect	Traditional Farming	IoT & AR-Based Agriculture
Decision Making	Experience-based	Data-driven
Monitoring	Manual	Real-time
Resource Use	Inefficient	Optimized
Technology	Minimal	Advanced
Productivity	Moderate	High
Labor	High	Reduced
Data Interpretation	Limited	Visual & interactive

F. Research Gap

Although IoT and AR technologies have been widely studied individually, research focusing on their integration in agriculture remains limited. Most existing studies emphasize IoT-based monitoring systems or AR visualization independently, with fewer addressing their combined application. Recent studies emphasize the need for better integration of IoT, AR, and AI to develop scalable and intelligent farming systems [7]. Additionally, challenges such as system interoperability, cost, and user adoption continue to hinder widespread implementation.

Additionally, several gaps remain in the implementation of IoT and AR technologies in agriculture. One of the major issues is the limited adoption among small-scale farmers, mainly due to financial and technical constraints. There is also a lack of cost-effective and scalable solutions that can be easily implemented across different farming environments and agricultural scales. Furthermore, the integration of these technologies with AI and predictive analytics is still insufficient, limiting the ability to perform advanced forecasting and intelligent decision-making. Usability and user experience challenges also continue to affect adoption, as some systems are complex and difficult for farmers to operate effectively without proper training and technical support. Here, it can be stated that IoT and AR technologies play a crucial role in converting agriculture into a more efficient and sustainable system.

Addressing these gaps is essential for advancing the practical implementation of IoT-AR systems in agriculture.

Despite the increasing interest in the application of IoT and AR in agriculture, the existing body of literature remains fragmented and lacks a comprehensive synthesis. Many studies focus on specific applications or technologies in isolation, resulting in a limited understanding of the overall research landscape. In such emerging and interdisciplinary domains, bibliometric analysis is particularly valuable, as it enables researchers to systematically analyze large volumes of scientific publications, identify research trends, and uncover knowledge gaps [10], [11]. Furthermore, bibliometric methods provide a structured approach to mapping the intellectual, conceptual, and social structure of a research field, which is often difficult to achieve through traditional literature reviews alone [10]. As highlighted in previous studies, bibliometric analysis is especially useful in areas where research is still developing or dispersed, as it offers a comprehensive overview and helps establish a foundation for future research directions [10], [12]. Therefore, given the limited and scattered literature on the integration of IoT and AR in agriculture, a bibliometric analysis is necessary to systematically evaluate existing research, identify emerging themes, and guide future studies.

III. METHODOLOGY

This section provides the systematic approach used to retrieve, filter, and analyze the bibliometric data related to IoT and AR in agriculture, limited to publications from 2020 to 2026. This analysis examined the publications collected from the Elsevier Scopus database based on documents categorized by conference papers, article papers, book chapters, and other related document types, as in Fig. 1.

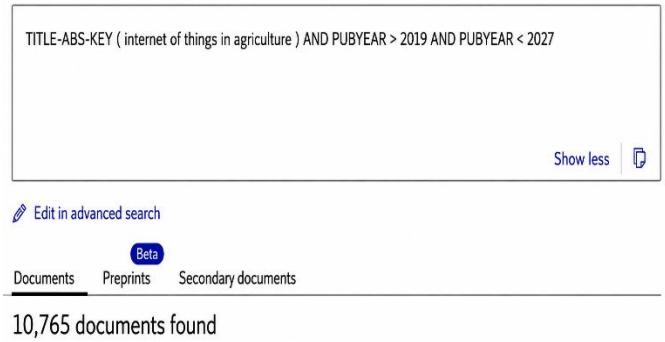


Fig. 1. Internet of Things (IoT) in agriculture.

Based on the search done in Scopus using the keyword Internet of Things and Augmented Reality in agriculture, there are around 72 related documents being generated. These numbers are quite low as compared to only one technology search. A search for IoT in agriculture from 2020 to 2026 had resulted in 10615 documents generated, as shown in Fig. 1 and a search for AR in agriculture from 2020 to 2026 had produced 280 documents, as in Fig. 2. Here it can be seen that the

integration of both technologies are still lacking in implementation and literatures, as in Fig. 3.

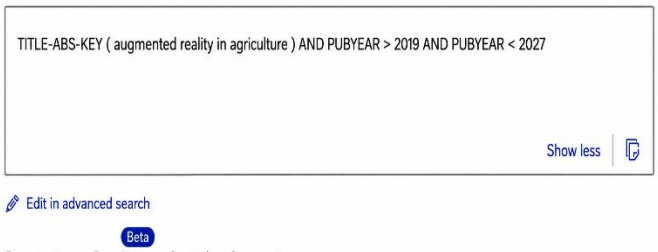


Fig. 2. Augmented Reality in agriculture.

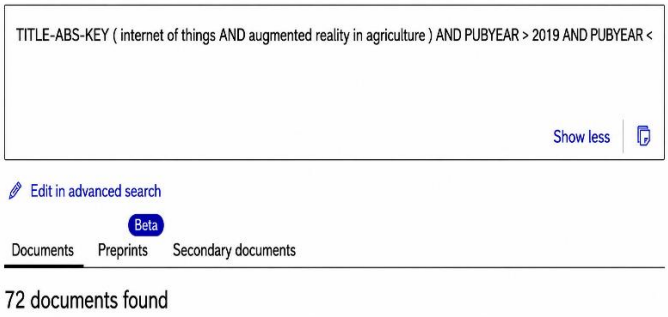


Fig. 3. Internet of Things and Augmented Reality in agriculture.

A detailed explanation of the search strategy is discussed in the following section. Fig. 4 below shows the distribution of all the 72 documents by year. The majority of the documents were published in 2024, with a total of 20 documents, followed by 2023 with 13 documents and in 2025 with 12 documents. Based on the graph, it can be seen that in 2026, the graph decreases with only 2 documents. This is due to the fact that during the search process, it is still in the third month (March) of 2026. There are still 9 months remaining, which will probably result in an increase in the number of documents later.

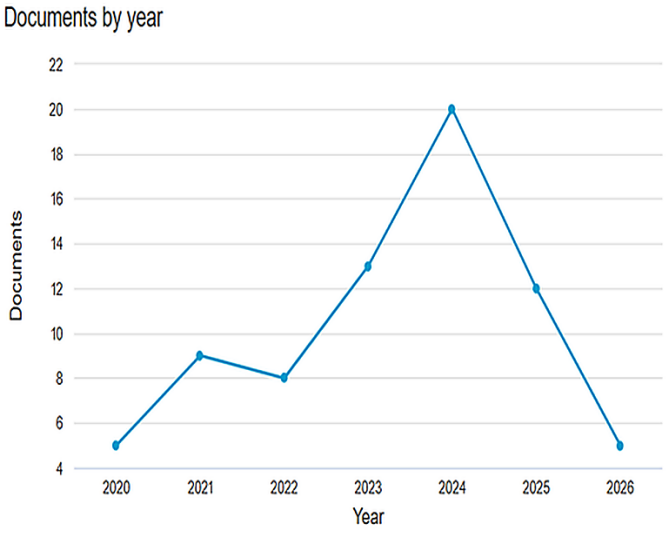


Fig. 4. Total number of documents by year.

A. Search Strategy

In order to retrieve relevant documents, a Boolean search string was developed that captured the intersection of data overload in real-time data processing. The query was applied to the title "Internet of Things and Augmented Reality in agriculture" with a filter of year from 2019 to 2026. Fig. 5 shows an illustration of a flowchart of the methodology used for conducting this bibliometric analysis using the Scopus database.

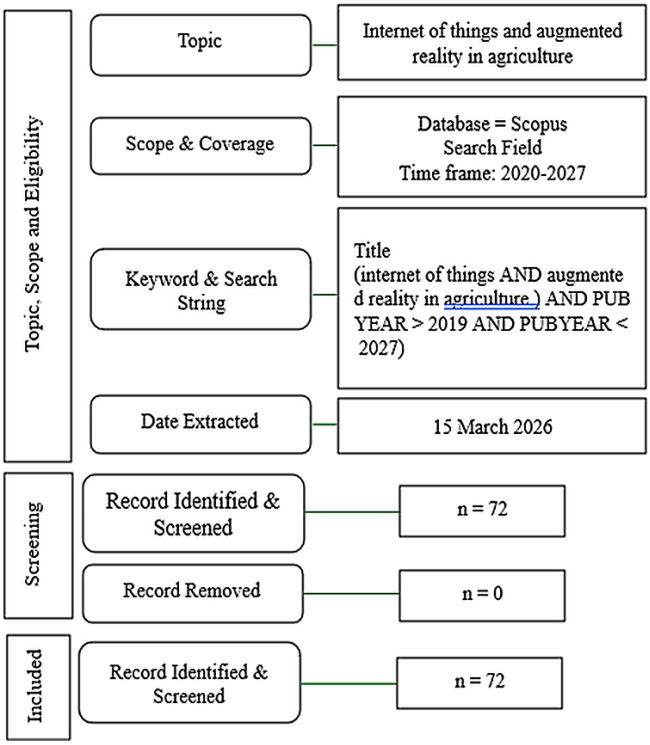


Fig. 5. Flow diagram of the strategy (Source: [13] as cited from [14]).

Fig. 5 consists of three main sections, which are: 1) Topic, Scope, and Eligibility, 2) Screening, and 3) Included. The first section outlines the foundational aspects of the research, where the topics used are stated, the scope and coverage of the study, the title of the search and also when the documents are extracted. It can be seen in the second section, Screening, where the total amount of documents was listed, which is 72 documents. In this search, there are no removed documents because the keywords are specifically filtered in the searching process on the Scopus database. The documents that have been searched are in detail already; therefore, there is no record removed and it indicates that all the records initially identified were considered relevant to the analysis. This analysis comprises three components listed below:

- The pattern in citations is examined to discover cited publications, influential authors, and related research groups.
- The themes and concepts on real-time data processing is analyzed based on co-occurring keywords.
- The emerging trends, areas and limitations of the area are discussed to offer a view of the research landscape and guidance.

This study offers insight into fundamental knowledge frameworks, research patterns, and trends of the technology used in visualizing real-time data.

B. Data Preparation Using VOSViewer

The Scopus-exported CSV file was converted to a format compatible with VOSviewer. The files were exported with metadata, as in Fig. 6:

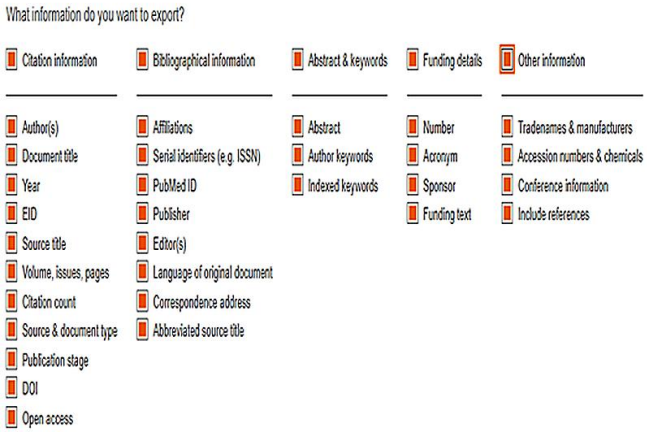


Fig. 6. Metadata information exported from the Scopus database.

Fig. 6 shows various categories of metadata fields commonly used to represent a different type of bibliographic information that helps identify, classify, and analyze research outputs. These categories are important for researchers, especially when conducting literature reviews, citation analysis, or bibliometric studies. The following figures show the basic settings of threshold, number of keywords, and the selected keywords.

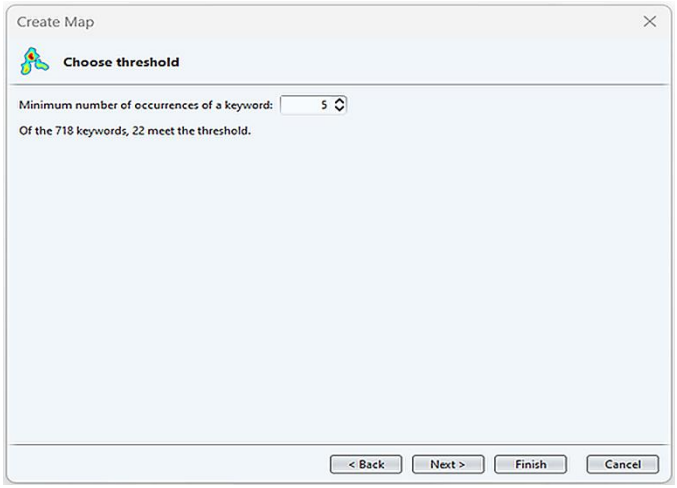


Fig. 7. Default maximum number of occurrences.

Fig. 7 shows the selection of the minimum number of keyword occurrences. Here, the value is set to 5 and based on this value, from the total of 718 keywords, only 22 keywords met the threshold. Here, it can be seen that only keywords that appear at least 5 times are considered important enough to represent the research. This step is important because it removes low-frequency or irrelevant keywords and also ensure that the analysis provides significant and frequently studied topic field.

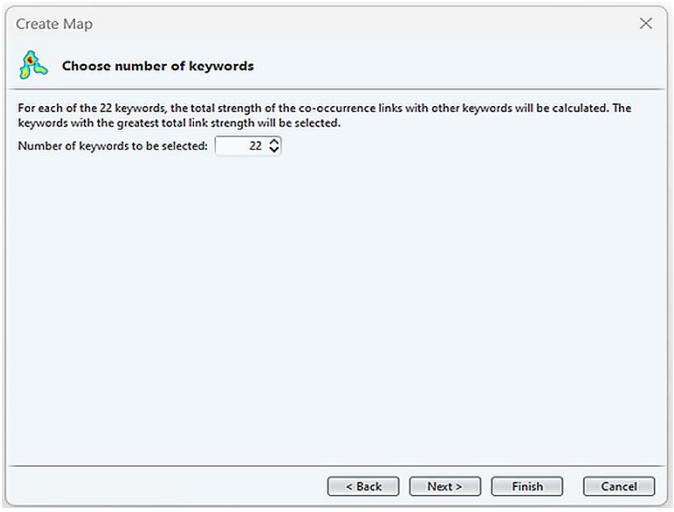
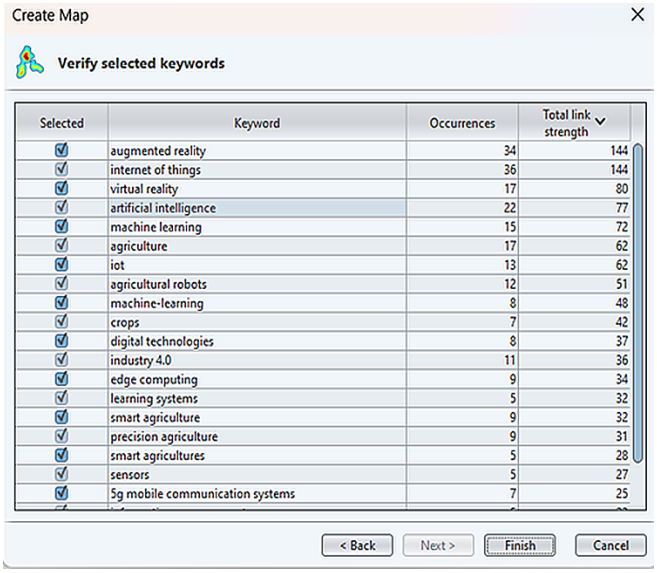


Fig. 8. Default number of keywords.

In Fig. 8, the selection of the 22 keywords is based on the total link strength. This total link strength measures how strongly a keyword is connected to other keywords and also measures the frequency of co-occurrence relationships of the keywords. Therefore, keywords with stronger connections represent core concepts and relationships within the specified research domains.



Selected	Keyword	Occurrences	Total link strength
☑	augmented reality	34	144
☑	internet of things	36	144
☑	virtual reality	17	80
☑	artificial intelligence	22	77
☑	machine learning	15	72
☑	agriculture	17	62
☑	iot	13	62
☑	agricultural robots	12	51
☑	machine-learning	8	48
☑	crops	7	42
☑	digital technologies	8	37
☑	industry 4.0	11	36
☑	edge computing	9	34
☑	learning systems	5	32
☑	smart agriculture	9	32
☑	precision agriculture	9	31
☑	smart agricultures	5	28
☑	sensors	5	27
☑	5g mobile communication systems	7	25

Fig. 9. Selected keywords.

Fig. 9 shows all of the keywords that are selected based on the setting in Fig. 7 and Fig. 8. Internet of things, augmented reality, and virtual reality are the three tops keywords with the highest occurrences and total link strength. Occurrences are defined as how often each keyword appears and total link strength is defined as how strongly it connects with others.

By applying a minimum occurrence threshold of five, the dataset was refined from 718 to 22 significant keywords, improving the clarity and reliability of the visualization.

The selected keywords demonstrate strong interconnections, with "Internet of Things" and "Augmented Reality" emerging as the most influential terms based on both occurrence and total link strength. This indicates that these technologies form the core of the research domain. Additionally, the presence of keywords, such as artificial intelligence, machine learning, and edge computing, highlights the interdisciplinary nature and evolving trends in smart agriculture research.

To explore the intellectual structure and research evolution of IoT and AR applications in agriculture, VOSviewer was employed to generate three complementary visualization types: Network Visualization, Overlay Visualization, and Density Visualization. Each visualization provides a different perspective on the relationships and development of research topics within the selected literature, which is discussed in the following section.

IV. Results

Keywords play an important role in bibliometric studies, as they reflect the main concepts, research interests, and thematic focus of a particular field. By examining the relationships among keywords, it is possible to identify the dominant research topics and understand how different concepts are connected within the literature. In this study, keyword co-occurrence analysis (Fig. 10) was conducted using VOSviewer to explore the research landscape of IoT and AR applications in agriculture. The resulting visualization provides insights into the major research themes, their interrelationships, and the emerging areas that have attracted increasing attention from researchers.

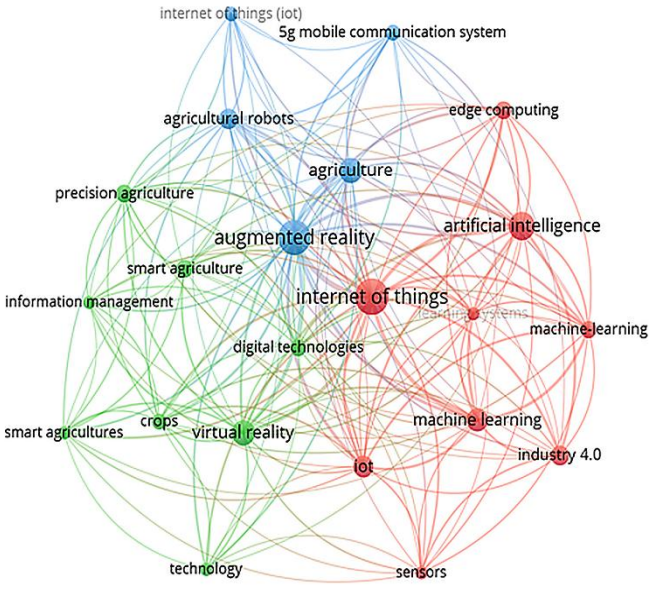


Fig. 10. Network visualization

The keyword co-occurrence network generated using VOSviewer reveals the conceptual structure of research on the integration of IoT and AR in agriculture. The analysis identifies several highly interconnected keywords, with "internet of things", "augmented reality", and "agriculture" emerging as the most prominent nodes, indicating their central role in the research domain.

The network is grouped into three primary clusters representing distinct but interconnected research themes. The first cluster focuses on smart agriculture and precision farming, incorporating keywords such as precision agriculture, smart agriculture, crops, and information management. This cluster highlights the application of IoT technologies in monitoring and optimizing agricultural processes.

The second cluster represents intelligent systems and data-driven technologies, including artificial intelligence, machine learning, sensors, and Industry 4.0. This cluster reflects the growing integration of advanced computational methods to enhance agricultural decision-making and automation.

The third cluster emphasizes technological infrastructure and communication systems, consisting of keywords such as 5G mobile communication systems, edge computing, and agricultural robots. This cluster illustrates the enabling technologies that support real-time data transmission and system interoperability. The strong interconnections among these clusters indicate that research in this field is highly interdisciplinary, combining sensing technologies, immersive visualization, artificial intelligence, and communication networks to support smart agriculture.

While the keyword co-occurrence network provides an overview of the major research themes and their relationships, it does not reveal how these themes have evolved over time. To address this limitation, an overlay visualization (Fig. 11) was generated using VOSviewer. This visualization builds upon the keyword network by incorporating temporal information, allowing researchers to identify both established and emerging research topics within the field.

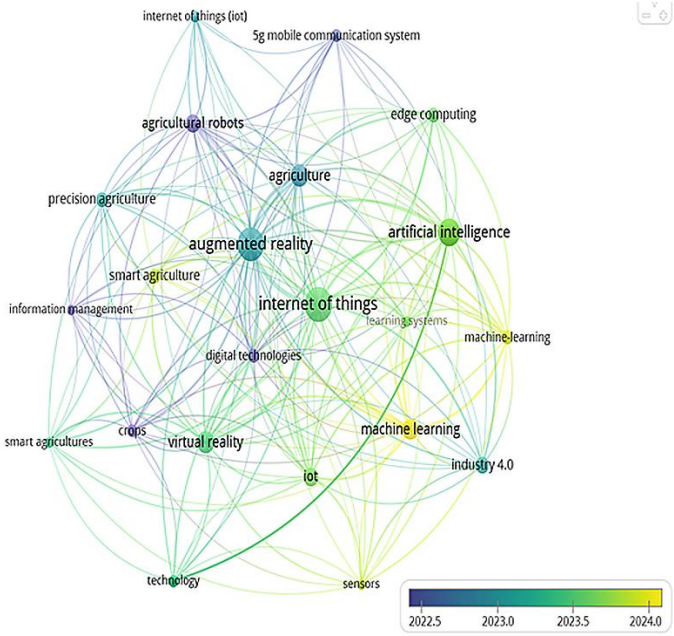


Fig. 11. Overlay visualization

The overlay visualization illustrates the temporal development of research topics in the field. Earlier studies (represented in blue tones) focused on foundational topics such as precision agriculture, information management, and

agricultural robots, reflecting the initial phase of digital transformation in agriculture.

Subsequent research (green tones) shows a shift toward augmented reality, smart agriculture, and digital technologies, highlighting the increasing importance of interactive visualization and user-centered systems.

More recent studies (yellow tones) emphasize artificial intelligence, machine learning, and Industry 4.0, indicating a transition toward intelligent, autonomous, and data-driven agricultural systems. This temporal progression suggests that the field has evolved from basic monitoring and automation to more advanced systems that integrate AI with IoT and AR for predictive and real-time decision-making.

Following the overlay visualization analysis, a density visualization (Fig. 12) was generated to further examine the level of research attention received by different topics within the field. While the overlay visualization highlights the temporal evolution of research themes, it does not indicate the relative concentration or prominence of individual topics. Therefore, density visualization is included as a complementary analysis to identify the most intensively studied areas and reveal the overall research hotspots within the literature.

Fig. 12. Density visualization

The density visualization provides insight into the intensity of research focus within the field. Areas highlighted in bright yellow indicate highly researched topics, while green and blue areas represent moderate and emerging topics, respectively. The analysis shows that “internet of things”, “augmented reality”, “agriculture”, and “artificial intelligence” are the most intensively studied topics. These keywords form the core foundation of current research and indicate strong academic interest in integrating sensing technologies with immersive visualization.

Moderate-density areas include machine learning, smart agriculture, and digital technologies, suggesting that these topics are gaining traction but are still developing. In contrast, keywords such as edge computing, Industry 4.0, virtual reality, and sensors appear in lower-density regions, indicating that these areas are relatively underexplored and present opportunities for future research.

V. Discussion

The findings of this bibliometric analysis demonstrate that the integration of IoT and AR in agriculture is rapidly evolving into a multidisciplinary research area. The dominance of core technologies such as IoT and AR reflects their critical role in enabling real-time data acquisition and immersive visualization.

The patterns identified in this study share several similarities with findings reported in earlier bibliometric research on smart farming, precision agriculture, and digital agriculture. Across these studies, IoT consistently appears as one of the core technologies driving agricultural digitalization. The keyword network generated in the present study reinforces this observation, with IoT emerging as one of the most influential and highly connected keywords. This suggests that IoT continues to serve as the technological foundation for collecting, transmitting, and managing agricultural data in modern farming environments.

Another notable observation is the growing presence of Artificial Intelligence (AI), Machine Learning (ML), and Industry 4.0-related concepts. Previous bibliometric studies have highlighted a gradual movement away from basic monitoring systems towards more intelligent and predictive agricultural solutions. A similar trend can be observed in the overlay visualization of this study, where AI and ML-related keywords appear among the more recent topics. This development indicates that researchers are increasingly interested in extracting meaningful insights from agricultural data rather than focusing solely on data collection.

While many studies in precision agriculture emphasize technologies such as IoT, remote sensing, drones, and artificial intelligence, the findings of this study reveal a growing interest in Augmented Reality (AR). The appearance of AR as a prominent keyword within the network suggests that researchers are beginning to explore new ways of presenting and interacting with agricultural information. As farming systems become increasingly data intensive, AR offers the potential to improve how users interpret and respond to real-time information by providing a more intuitive and interactive visualization environment.

The density visualization further supports this observation. Keywords related to IoT, AR, AI, and machine learning were concentrated within the most active areas of the map, indicating that these topics currently attract significant research attention. This finding is consistent with broader discussions in the digital agriculture literature, where the integration of sensing technologies, intelligent analytics, and advanced visualization tools is often described as an important step towards more efficient and sustainable farming practices.

Although previous bibliometric studies have provided valuable insights into smart agriculture as a whole, relatively few have specifically examined the relationship between IoT and AR in agricultural applications. In this regard, the present

study contributes a more focused perspective by highlighting how immersive technologies are increasingly being combined with IoT-based systems. The findings suggest that AR is evolving beyond a supporting visualization tool and is becoming an important component of next-generation agricultural solutions. This creates opportunities for future research on real-time visualization, decision support, and user interaction within smart farming environments.

Furthermore, the presence of enabling technologies such as 5G and edge computing highlights the importance of communication infrastructure in supporting real-time agricultural applications. Despite these advancements, several areas remain underexplored, particularly the integration of virtual reality, advanced sensor networks, and scalable solutions for smallholder farmers. Addressing these gaps will be essential for the widespread adoption of smart agriculture technologies.

VI. Research Challenges and Future Directions

The growing number of publications identified in this study reflects the increasing interest in applying IoT and AR technologies within the agricultural sector. However, despite the promising developments reported in the literature, several challenges continue to limit the widespread adoption of these technologies in real farming environments.

One of the most frequently cited concerns is the cost of implementation. Although IoT sensors, cloud-based platforms, and AR applications can provide valuable insights for farm management, the initial investment required for hardware, software, and system maintenance may be beyond the reach of many small-scale farmers. As a result, the adoption of advanced digital farming technologies is often more feasible for larger agricultural enterprises with greater financial resources.

In addition to financial considerations, the successful use of IoT and AR technologies depends on users' ability to interact with digital systems and interpret the information provided. Farmers may face difficulties when dealing with unfamiliar technologies, particularly when systems are complex or require specialized technical knowledge. This highlights the importance of designing user-friendly solutions that simplify data interpretation and support practical decision-making in agricultural activities.

Infrastructure availability also plays a significant role in technology adoption. Many IoT-based applications rely on stable internet connectivity, reliable power sources, and access to cloud services. In rural and remote farming areas, these requirements may not always be fully available, creating barriers to implementation. Although emerging technologies such as edge computing and next-generation communication networks offer potential solutions, their accessibility remains uneven across different regions.

Beyond technical and economic factors, policy support is another important consideration. Government initiatives, funding schemes, and digital agriculture programs can help reduce adoption barriers by providing financial assistance, training opportunities, and infrastructure development. Without adequate policy support, the benefits of IoT and AR technologies may remain limited to a relatively small group of technology-ready users.

The findings of this study suggest that future research should extend beyond technological innovation alone. Greater attention should be given to understanding how economic constraints, user readiness, infrastructure availability, and policy support influence the adoption of smart agriculture technologies. Addressing these factors will be essential to ensure that the benefits of IoT and AR can be accessed by a wider range of agricultural stakeholders and contribute meaningfully to sustainable agricultural development.

VII. Conclusion

This study presents a comprehensive bibliometric analysis of research on the integration of Internet of Things (IoT) and Augmented Reality (AR) in agriculture based on Scopus-indexed publications from 2020 to 2026. The findings indicate a growing research interest in this interdisciplinary domain, with a notable increase in publications, particularly in recent years. This trend reflects the increasing importance of digital technologies in transforming traditional agricultural practices into more efficient and sustainable systems.

The VOSviewer analysis provides deeper insights into the intellectual structure and evolution of the research field. The keyword co-occurrence network reveals that Internet of Things, Augmented Reality, and agriculture serve as the central themes, supported by closely related concepts such as precision agriculture, artificial intelligence, and smart farming systems.

The clustering structure highlights the interdisciplinary nature of the field, integrating sensing technologies, data analytics, communication systems, and immersive visualization. The density visualization further confirms that IoT and AR are the dominant research hotspots, while emerging topics such as edge computing, virtual reality, and Industry 4.0 remain less explored. This suggests that while the core technologies are well established, there are still significant opportunities for expanding research into advanced and integrated systems.

In addition, the overlay visualization demonstrates a clear temporal evolution of the field. Early research primarily focused on data collection and monitoring through IoT, followed by the incorporation of AR for visualization and interaction. More recent studies emphasize the integration of artificial intelligence and machine learning, indicating a shift toward intelligent, predictive, and autonomous agricultural systems. This progression reflects the movement from basic digital agriculture toward fully integrated smart farming ecosystems.

The findings from the literature review further support this evolution, highlighting the complementary roles of IoT and AR in enabling real-time data acquisition and intuitive data visualization. While IoT enhances monitoring and automation, AR improves data interpretation and user interaction, thereby bridging the gap between complex data and practical decision-making. However, challenges such as high implementation costs, infrastructure limitations, and usability issues continue to hinder widespread adoption.

Overall, this study confirms that the integration of IoT and AR has significant potential to revolutionize agriculture by improving productivity, efficiency, and sustainability. Nevertheless, the bibliometric analysis also reveals that the research landscape remains fragmented, with limited studies

focusing on the combined application of these technologies. This highlights the need for more comprehensive and integrated research efforts.

Future studies should focus on the development of scalable and cost-effective solutions, particularly for smallholder farmers, as well as the integration of emerging technologies such as artificial intelligence, digital twins, and advanced communication systems. In addition, further research is needed to improve system usability, enhance interoperability, and address practical implementation challenges.

In conclusion, this study provides a structured and comprehensive overview of the research landscape on IoT and AR in agriculture. It contributes to the existing body of knowledge by identifying key trends, themes, and research gaps, and offers valuable insights to guide future research toward the development of intelligent and sustainable agricultural systems.

ACKNOWLEDGMENT

This research was supported and fully funded by the Ministry of Higher Education (MOHE) Malaysia through the Fundamental Research Grant Scheme – Early Career Researcher (FRGS-EC/1/2024/ICT09/UUM/03/1). The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of MoHE, Malaysia. The authors acknowledge the use of AI-based tools, including ChatGPT, for language support and structural refinement. The use of this tool was conducted under strict human supervision to ensure the accuracy, clarity, and integrity of the research. All ideas, analyses, interpretations, and final decisions remain the full responsibility of the authors.

REFERENCES

[1] N. S. Abu, W. M. Bukhari, C. H. Ong, A. M. Kassim, T. A. Izzudin, M. N. Sukhaimie, M. A. Norasikin, A. F. A. Rasid, "Internet of Things applications in precision agriculture: A review," Journal of Robotics and Control (JRC), 3 (3), 338-347, 2022.

[2] S. Mansoor, S. Iqbal, S.M. Popescu, S. L. Kim, Y. S. Chung and J. H. Baek, "Integration of smart sensors and IoT in precision agriculture," *Frontiers in Plant Science*, 2025.

[3] M. Raj and M. Prahadeeswaran, "Revolutionizing agriculture: a review of smart farming technologies for a sustainable future" in *Discover Applied Sciences*, *7*(9), 937.

[4] A. Subeesh, N. Chauhan "Agricultural digital twin for smart farming: A review," in Green Technologies and Sustainability, 100299, 2025.

[5] T. Y. Melesse, "Digital twin-based applications in crop monitoring," in Heliyon, 2025.

[6] T. Miller, G. Mikiciuk, I. Durlik, M. Mikiciuk, A. Lobodzinska and M. Snieg, "The IoT and AI in agriculture: The time is now - A systematic review," *Sensors*, vol. 25, no. 12, 3583 2025.

[7] A. Cissé, O. Diallo, E. H. M. Ndoye, O. Sene, "Review of smart agriculture for resource-limited areas: IoT and AI integration," *Precision Agriculture*, *Discover Computing*, 28(1), 324 2025.

[8] D. Muhammed, E. Ahvar, S. Ahvar, M. Trocan, M. J. Montpetit and R. Ehsani "Artificial Intelligence of Things (AIoT) for smart agriculture: A review of architectures, technologies and solutions," *Journal of Network and Computer Applications*, 228, 103905, 2024.

[9] M. M. Hussain and M. E. Islam, "A systematic literature review on IoT integrated agriculture: Security and productivity," *International Journal of Research and Scientific Innovation*, vol. 13, no. 13, pp. 30–48, 2026.

[10] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, vol. 133, pp. 285–296, 2021.

[11] I. Passas, "Bibliometric analysis: The main steps," *Encyclopedia*, vol. 4, no. 2, 2024.

[12] S. Greener, "Evaluating literature with bibliometrics," *Interactive Learning Environments*, 30(7), 1168-1169, 2022

[13] N. S. Ishak,, M. F. Roslan and K. H. Abdullah, "Bibliometric analysis of Malaysian authorship: Trends, patterns, and prospects". *Asian Journal of Research in Education and Social Sciences, 5*(2), 33–44.

[14] Zakaria, N., Abdullah, N. A. C., Zahoor, N., Azizan, M., & Syed, Z. A. A. (2024). A bibliometric analysis of quality of work-life: Current status, development, and future research directions. *Pakistan Journal of Life and Social Sciences, 22*(2), 326–347.

Evolutionary Allocation with Dynamic Guidance for Pre-Scheduled Timetables

Anniza Hamdan[1], Sze San Nah[2], Goh Say Leng[3], Emily Sing Kiang Siew[4]
Faculty of Computer Sciences and Mathematics,
MARA University of Technology, Kota Samarahan, Sarawak 94300, Malaysia[1]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,
Kota Samarahan, Sarawak 94300, Malaysia[2]
Optimization and Visual Analytics Research Group-Faculty of Computing and Informatics,
Universiti Malaysia Sabah, Labuan International Campus, Labuan 87000, Malaysia[3]
i-CATS University College, Kuching, Sarawak 93350, Malaysia[4]

***Abstract*—The University Course Timetabling Problem (UCTP) is a well-known combinatorial optimization challenge that involves allocating courses to timeslots and rooms while satisfying various institutional constraints. At other institutions, general courses (e.g., language subjects) are prioritised during timetable allocation due to their high enrolment numbers from multiple faculties. However, at the MARA University of Technology, Sarawak Branch, Mathematics and Statistics (MAT/STA) courses are shared across several programs, with timeslot availability limited by pre-scheduled major courses in each program. This study presents a tailored evolutionary algorithm for a real institutional scenario, incorporating dynamic local search and guided variation operators to improve feasibility and solution quality. A case study was conducted using real datasets from the Department of Mathematical Sciences at MARA University of Technology, Sarawak Branch. Benchmark datasets from ITC2002 and ITC2007 (Track 2) are employed for comparative evaluation against existing methods. The algorithm successfully produced a feasible timetable and outperformed manually prepared schedules in terms of soft-constraint penalties. Results indicate strong performance on real datasets, producing a high-quality, balanced timetable with reduced preparation time, particularly in handling repeating students, lecturers' availability, and limited timeslots. Nevertheless, lower performance on benchmark datasets suggests the need for hybridization or additional operators to handle larger, more complex problems. Overall, the results demonstrate the effectiveness of the proposed approach in real-world timetabling while maintaining acceptable quality on benchmark instances.**

Keywords*—Timetabling problem; evolutionary algorithm; dynamic; guided

I. Introduction

The UCTP is a well-known NP-hard optimization challenge that involves assigning courses, rooms, timeslots, and student groups under various hard and soft constraints [1], [2], [3]. The constraints of UCTP vary across academic institutions due to their respective policies and regulations. Practically, the goal of UCTP is to provide a balanced timetable with minimal conflicts. Hard conflicts may cause an infeasible timetable, such as when a lecturer or students have two courses scheduled at the same time. Soft conflicts determine the quality of the timetable, such as a morning or afternoon class overload, and too many consecutive lecture hours in a day.

This research is motivated by a different practice of course timetabling in MARA University of Technology (UiTM), Sarawak Branch. The Department of Mathematical Sciences at UiTM Sarawak Branch offers MAT/STA courses across 14 programs, including Public Administration, Accountancy, Engineering, and more. Current timetabling prioritizes the allocation of major courses (of programs). Each major course may have multiple weekly sessions, for example, 2 hours on Day 2, 2 hours on Day 3, and 1 hour on Day 4 (2+2+1). The pattern of these multiple weekly sessions can be 3+0, 2+2, or 3+1. Consequently, the available slots of 2+2 MAT/STA courses are limited. In addition, each lecturer must satisfy minimum teaching hours and shared teaching requirements. At the same time, parallel timeslots for MAT/STA courses are necessary to accommodate students who repeat across multiple program groups.

Timetable preparation begins after the final examination of each semester. The projected numbers of repeat students and enrolments are used to estimate the required groups for each MAT/STA course. A program-level group projection (may take 1-2 weeks) must be completed first. Once receiving these projections, the Department of Mathematical Sciences may further adjust the number of required groups for each MAT/STA course based on teaching load and shared-teaching requirements for the semester. The existing timetabling system allows users to view available slots and manually assign MAT/STA courses to program or group timetables. Nonetheless, since multiple programs take these courses, timetable conflicts must be checked manually. Some lecturers may request fewer consecutive teaching hours in a day, while others may express concerns about the balance between morning and afternoon teaching slots. Repeat students are also facing issues with their repeated courses. These adjustments and checks continue until the semester starts (commonly stable after week 4 of lectures).

Furthermore, the allocation of the MAT/STA course can adversely affect timetable quality by creating an imbalance in daily load. For instance, a pre-scheduled timetable already contains six lecture hours on days 1, 2, and 3. Assigning additional sessions to these days increases the consecutive hours per day. Redistributing the MAT/STA course to the lighter days 4 and 5 would promote better load balancing and enhance overall timetable efficiency. Once the semester begins, further

adjustments may compromise the stability and feasibility of the established timetable. Therefore, an efficient strategy is required to address this issue.

Motivated by real-life applications, the problem aims to minimize the conflicts while providing a balanced timetable in a short time. Three semesters of data are collected. An evolutionary algorithm (EA) is proposed to explore possibilities for MAT/STA course allocations while minimizing conflicts and achieving a balanced timetable. There are previous studies that have been done for real-life applications using EA variants: Genetic algorithm [4], [5]; Particle Swarm Optimization [1], [2], [3], [6]; and Hybrid Evolutionary Algorithm [7], [8], [9]. Most of these studies introduce new operators (crossover and mutation) during the evolutionary stage to address the nature of real-world applications. Some operators allow temporary conflicts during evolution and refine them later. Hybridization with local search was performed to refine the solution. Hence, the proposed algorithm incorporates a dynamic local operator and guided variation operators to refine the search process and improve timetable quality adaptively.

II. Formulation and Approach

A timetable is represented as a three-dimensional (3D) array (groups/lecturer, day, and timeslot). A binary decision variable, $X_{c,t,r}$, is used (which takes the value 1) when a lecture of the course c is allocated to a day d and timeslot t. Given that $R_{t,r} = 1$ indicates that a room is counted for each timeslot t. All MAT/STA courses for the assigned groups must be scheduled each week into available timeslots and rooms. Multiple weekly sessions of 2+2 MAT/STA lecture hours are fully allocated, and each lecturer teaches only one MAT/STA course per timeslot and room. Sufficient rooms are available for all multiple weekly sessions, and no MAT/STA courses are assigned to blocked timeslots. Blocked timeslots are defined as follows:

- Last timeslot of each day, $t = \{9,18,27,36,45\}$.
- Lunch hour, 1 p.m. until 2 p.m. (not included in a set of timeslots).
- Extra co-curricular activity, Wednesday 4 p.m. until 6 p.m. for year 1 students (not included in a set of timeslots).
- Friday prayer, Friday 12 noon until 2 p.m. (not included in a set of timeslots).

To achieve a balanced, high-quality timetable, MAT/STA courses with a 2+2 lecture hours should be evenly distributed throughout the week, with minimal preset gaps. The students' group should not have more than 7 consecutive lecture hours per day, and isolated 2-hour MAT/STA lectures should be avoided. Penalties are applied whenever these preferences are violated.

To the best of our knowledge, there is no real-world application of previous studies using an evolutionary algorithm that considers a session gap in course allocation with remaining timeslots (major courses of a program are scheduled first). Maintaining the crucial session gap while allocating courses to limited time slots complicates the standard UCTP. The most relevant paper [9] used an improved genetic algorithm on the benchmark dataset ITC2007 (Track 2). However, the allocation of events in ITC2007 (Track 2) is based solely on available timeslots, without any specific session gaps between events.

An EA is proposed in this study due to its flexibility in handling diverse constraints of the problem. It allows interaction and information exchange in a population-based environment [10]. The main steps of EA consist of the following: 1) Initialization, 2) Fitness evaluation, 3) Parents selection for reproduction, 4) Variation operators, 5) Selection of the best solution [11].

As the case study includes an additional hard constraint of a 2+2 session gap for each MAT/STA course, a dynamic local operator is specifically designed and applied during initialization and variation operators. Excluding this operator will produce infeasible timetables. A steady-state algorithm generates a population size of [20,100] at the initial stage. The improvement stage evolves this population iteratively using variation operators, generating a complete, updated solution.

A. Initialization

In EA, random initialization may lead to an infeasible solution when constraints are not enforced. Rezaeipanah et al. [9] applied a random shuffling technique to promote genetic diversity and population diversity. Nevertheless, in this study, we apply sorting and shuffling techniques to ensure that the high-constraint student groups are prioritized.

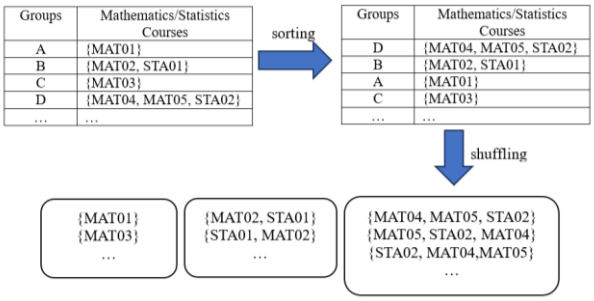


Fig. 1. Sorting and shuffling technique in the initialization phase.

Fig. 1 illustrates the MAT/STA courses ranked by the number of courses taken by each group. The MAT/STA courses in this sorted list are shuffled to produce different initial allocations. The iteration continues until the desired population size is achieved.

If any conflicts are found, a dynamic local operator is used to adjust the preset session gap for MAT/STA courses; it is necessary to obtain a feasible initial solution. The adjustment is made by reducing the preset session gap value for the course, which may result in a session gap penalty later on. Nonetheless, due to the highly constrained MAT/STA groups, some runs may not achieve the maximum population size.

B. Fitness Evaluation

Let $X_i, i = 1,2,3, \dots P$, denote a candidate timetable in a population P. The solution quality is measured based on a weighted soft-constraint penalty objective function:

$$F_{obj}(X) = w_1 \sum_{c \in C} Q_c + w_2 \sum_{g \in G} \sum_{d=1}^{5} C_{g,d} + w_3 \sum_{g \in G} \sum_{d=1}^{5} I_{g,d.} \quad (1)$$

In Eq. (1), Q_c measures session-gap violations for the course c. $C_{g,d}$ denotes excess consecutive hours for the group g on day d, and $I_{g,d}$ counts any isolated sessions. The coefficients w_1, w_2, w_3 are fixed values of penalty weights for each soft constraint violation. Each soft-constraint violation type employs a predefined penalty weight with a predefined cost: $w_1 = 1$ for session-gap violations, $w_2 = 7$ for excessive consecutive hours, and $w_3 = 1$ for an isolated MAT/STA course within a day. The objective is to minimize $F_{obj}(X)$, where smaller values indicate higher timetable quality.

One advantage of a sum-weighted penalty function is that it can be applied to any optimization technique, as it focuses on minimizing a single objective function. In addition, it is flexible and easy to use, as the weight can be assigned based on its importance [12]. In the existing literature [13], [14], [15], the cost penalty for 3 consecutive hours is set to 1. Meanwhile, we significantly increased the value of w_2 to highlight the severity of pre-scheduled timetables. For the real datasets, the counting of consecutive hours also accounted for pre-scheduled slots on the same day. Consequently, allocating the MAT/STA course to the days that already contain long consecutive hours could negatively affect the overall quality and balance of the timetable.

C. Parents Selection

A k-tournament selection is used at this stage with $k = 3$. Three individual timetables are selected randomly and compared (fitness value). The timetable with the lowest penalty value is chosen as Parent 1. The process is repeated to select Parent 2, provided that Parent 2 is not the same individual as Parent 1. Fig. 2 illustrates the selection process for the Parent.

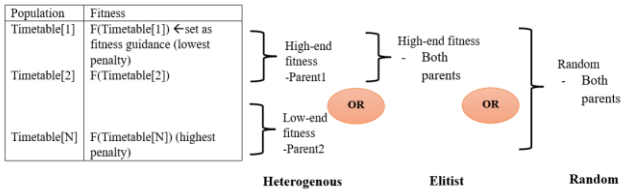


Fig. 2. Illustration of parent selection mechanisms used in the evolutionary process.

The proposed algorithm employs three types of selection strategies: 1) Heterogeneous, 2) Elitist, and 3) Random. Individuals in the population are ranked before selection. In Heterogeneous selection, Parent 1 is randomly selected from the high-end fitness individuals, and the other is randomly selected from the low-end fitness individuals to preserve diversity and balance exploration and exploitation. In Elitist selection, both parents are randomly selected from the high-end fitness individuals. Meanwhile, in Random selection, parents are selected randomly from the population.

D. Swapping 1 (External)

Previous studies have applied a single-point, uniform heuristic crossover and other variation techniques, such as swap sequence, forceful swap, switching tolerance, horizontal, and vertical swap, to generate offspring. Similarly, Swapping 1 purposely exchanges two candidate solutions (Parent 1: x_i and Parent 2: x_p) by matching a random event with the same MAT/STA course. This guided approach reduces randomness and complexity caused by limited timeslots, while same-course swapping helps avoid conflicts.

In candidate x_i, Session 1 of course c_i, scheduled on day 2 in timeslots 3-4, is swapped with Session 1 of the same course code in the timetable x_p, scheduled on day 1 in timeslots 1-2. However, Session 2 cannot be swapped directly because it conflicts with major courses. A dynamic local operator is activated to handle this conflict. Two offspring (x'_i and x'_p) are obtained. Fig. 3 demonstrates the external swapping between Parent 1 and Parent 2.

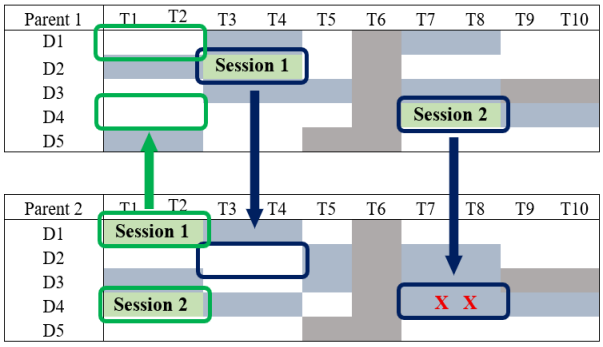


Fig. 3. External swap of two parent solutions, where selected sessions are exchanged. the dynamic local operator handles the resulting conflict ("XX").

E. Swapping 2 (Internal)

Swapping 2 functions like a mutation operator by performing a guided internal swap in each offspring (x'_i, x'_p) generated from Swapping 1. For each offspring, a random 2-hour session with the same MAT/STA course is selected. Session 1 in x'_{i1} (day 1, timeslots 3-4) is swapped with Session 1 in x'_{i2} (day 2, timeslots 9-10). As the timeslots in x'_{i2} are unavailable, the dynamic local operator reallocates the swapped Session 2 to the next available timeslots. Penalties are recalculated, and the offspring replaces the parent if it improves the solution. Fig. 4 illustrates the internal swapping in each offspring.

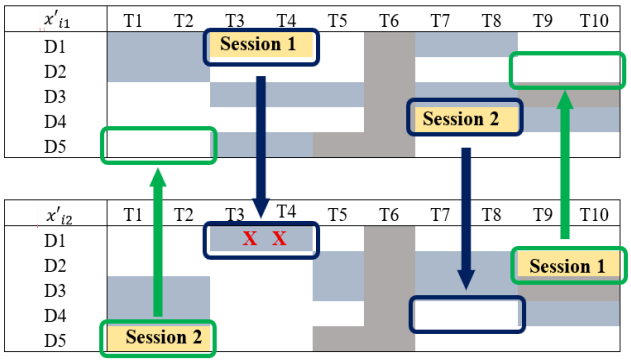


Fig. 4. Internal session swap within each offspring solution, where selected sessions are exchanged, and the resulting conflict ("xx") is repaired using the dynamic local operator.

F. Checking Termination Criteria

The algorithm is terminated when a maximum number of generations is reached, or no improvement is observed over a predefined number of consecutive generations.

G. Selection of the Best Solution

After the termination criteria are satisfied, the candidate solutions are ranked by their penalty values (objective function costs). The solution with the lowest penalty value is chosen as the final solution. The average penalty value is calculated from the solutions of the last generation.

H. Guided Variation Operator

In other institutions, general courses such as language subjects, which are taken by all programs, are prioritized during timetable allocation due to their complexity. In contrast, at the UiTM Sarawak Branch, a MAT/STA course is offered across multiple programs, with each program comprising several groups enrolled in the same MAT/STA course. Given this scenario, the same MAT/STA course was chosen for swapping to minimize the complexity of finding feasible timeslots, as variations in pre-scheduled major courses across groups increase timetabling complexity.

TABLE I. NUMBER OF GROUPS IN EACH PROGRAM TAKING THE MAT/STA COURSE.

Course Code	Program Code	Student Year	Number of groups		
			October 2022	March 2023	October 2023
Diploma Programs					
MAT111	BA132	1	5	4	7
MAT112	AC110	1	6	1	5
	AM110	1	7	5	7
	IC110	1	4	2	4
MAT133	AS125	1	1	1	1
	AS126	1	2	1	2
	AS120	1	6	2	4
	CS110	1	7	5	7
MAT183	EH110	1	4	0	2
	EC110	1	3	1	3
	EE111	1	0	0	1
	AS125	2	0	1	0
	AS120	2	0	7	1
	CS110	2	1	8	4
MAT210	CS110	3	6	3	8
MAT235	EC110	1	0	2	1
	EE111	1	0	1	0
	EH110	1	0	2	1
MAT238	AS120	3	4	1	1
MAT285	EC110	3	4	1	1
STA104	AC110	2	1	9	1
STA108	AS115/ AS120	3	2	6	2
STA116	CS110	3	2	6	2
Degree Programs					
MAT402	AM228	1	2	1	2
MAT421	AS222	1	1	0	1
STA404	AM228	2	5	6	5
	AS222	3	1	0	0
	AC220	3	1	2	1
STA408	AT220	2	1	3	1
	AS222	1	0	0	1
ITS665	AS222	3	1	0	1

Table I indicates the number of groups taking one or more MAT/STA courses. The difficulty of identifying feasible timeslots increases as more programs enrol in the MAT/STA course or when a program takes multiple MAT/STA courses (see Fig. 5).

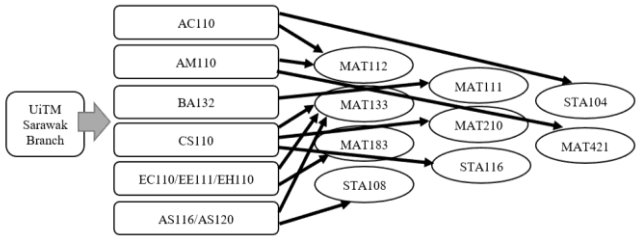


Fig. 5. Complexity of MAT/STA courses taken by multiple programs.

I. Dynamic Local Operator

This operator is crucial in Initialization, Swapping 1, and Swapping 2. While variation operators generate offspring to explore the solution space, this operator acts as a dynamic solver that resolves conflicts arising from guided variations, particularly in handling the session gaps. In the Initialization stage, a preset session gap is used to schedule a second session of a 2-hour MAT/STA course after the first one is successfully allocated. However, pre-scheduled major courses may occupy certain time slots, creating conflicts. Thus, Session 2 is relocated to the next available timeslot, and the session gap is recalculated. To prevent sessions from being scheduled too closely together, a penalty is applied when the updated gap is smaller than the preset session gap. Fig. 6 illustrates the relocation process for Session 2 (with a preset session gap of 18 timeslots).

The same process is applied in Swapping 1 and Swapping 2. Nevertheless, conflicts encountered during these swapping operations may necessitate relocating either or both of Sessions 1 and 2. This adaptive strategy ensures feasibility while permitting controlled relaxation to enhance exploration and minimise unallocated courses.

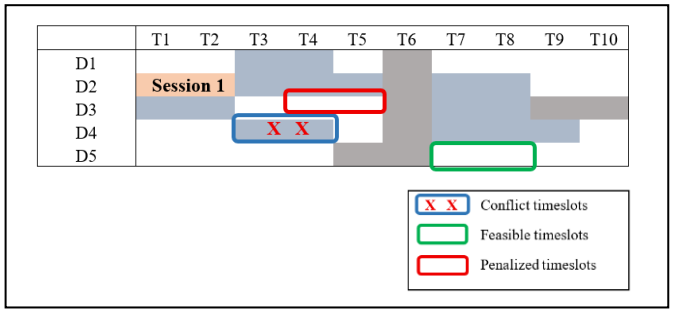


Fig. 6. Relocation process in the initialization stage.

III. EXPERIMENTS

A. Experimental Setting

In this case study, a real-world dataset is derived from an actual university timetable from the UiTM, Sarawak Branch. A real challenge and diverse policies across organizations in this dataset are studied to tailor the proposed algorithm and evaluate its performance. The details of the data are summarized in the following Table II:

TABLE II. CHARACTERISTICS OF REAL DATASETS FOR THREE SEMESTERS.

Semester	October 2022	March 2023	October 2023
No. of programs	15	15	15
No. of students	1637	1372	1408
No. of groups	78	67	81
No. of MAT/STA courses	14	14	16
No. of groups taking multiple MAT/STA courses	1(2 courses)	11(2 courses)	5(2 courses) 1(3 courses)
No. of lecturers	21	23	23
No.of MAT/STA courses with shared-teaching	0	7	3

The experiment was coded in Java, and all experiments were carried out on an Acer computer equipped with a 12th Gen Intel® Core™ i5-1235U processor running at 3.45 GHz. We performed 10 independent runs for each dataset. In addition, the ITC2002 and ITC2007 (Track 2) benchmark datasets were used as a baseline to evaluate the proposed algorithm's performance relative to existing methods.

In preliminary experiments, various settings were used to determine the best session gaps and end timetable hours for each dataset. The details of the setting tested are as follows:

- Session gaps: {9, 10, 11, 12, 13, 14, 15, 16, 17, 18}
- End timetable: {17, 18, 19, 20}

Preliminary experiments show that the best predefined session gaps for each semester are 11, 13, and 13, respectively. Nonetheless, the session gap for each dataset may vary for multiple MAT/STA courses due to adjustments made by the dynamic local operator.

B. Experimental Results

Table III indicates that penalty values (average of soft-constraint violations) were reduced by 56.19% in October 2022, 66.30% in March 2023, and 59.63% in October 2023, showing substantial improvements in lecture distribution, consecutive hours, and isolated sessions. Although some datasets required later end times and more timeslots, this trade-off enabled a more balanced distribution of parallel sessions and lower penalties in a more complex environment. Overall, the results demonstrate the method's effectiveness in generating a feasible and balanced timetable while addressing critical session gaps.

The algorithm performance was further evaluated in terms of the penalty cost achieved in the initial and improvement stages (see Table IV). Random initial solutions may take time to improve during an evolutionary stage [16]. A sorting and shuffling technique was applied in the initial stage which indicates the effectiveness of obtaining good-quality feasible solutions. The proposed guided variation operator also speeds up the improvement of each solution. The significant improvement is evident from the large deviation in penalty cost reduction during the improvement stage.

TABLE III. OVERALL PERFORMANCE (AVERAGE COST) OF THREE DATASETS.

Semester	October, 2022		March, 2023		October, 2023	
	Act.	Imp.	Act.	Imp.	Act.	Imp.
End timetable	18:00	18:00	17:00	18:00	17:00	19:00
Penalty Value	105	46 **↓56.19%**	181	61 **↓66.30%**	218	88 **↓59.63%**
Timeslots	46	46	42	46	42	50

Act. = Actual; Imp. = Improved

TABLE IV. STATISTICAL ANALYSIS FOR THE INITIAL STAGE AND IMPROVEMENT STAGE ACROSS THREE DATASETS.

Datasets	Stage	Minimum	Mean	Std. Deviation
October, 2022	Initial	21	52.16	10.88
	Improvement	0	46	**20.93**
March, 2023	Initial	49	68.54	8.17
	Improvement	0	61	**21.86**
October, 2023	Initial	98	143.32	13.71
	Improvement	7	88	**33.58**

A one-way analysis of variance (ANOVA) based on relative improvement was conducted, as shown in Table V, to evaluate the algorithm's stability across independent runs. The results indicate consistent algorithm performance across all datasets and multiple runs (p-value > 0.001).

TABLE V. ANOVA FOR EACH DATASET OF THE ALGORITHM

Datasets	Source	DF	Sum of squares	Mean Square	F	Signature
October, 2022	Between groups	9	5515.022	612.780	0.690	**0.718**
	Within groups	347	308235.712	888.287		
	Total	356	313750.734			
March, 2023	Between groups	9	1066.636	118.515	0.171	**0.997**
	Within groups	480	332711.768	693.150		
	Total	489	333778.403			
October, 2023	Between groups	9	951.823	105.758	0.264	**0.984**
	Within groups	463	185482.213	400.610		
	Total	472	186434.036			

TABLE VI. SESSION GAP AND CONSECUTIVE HOURS (IN A DAY) FOR ALL MAT/STA COURSES.

Semesters	October, 2022		March, 2023		October, 2023	
Stages	Act.	Imp.	Act.	Imp.	Act.	Imp.
Minimum Session Gap	6	9	3	9	4	19
Average Session Gap	18	13	18	22	10	21
Average Consecutive hours in a day	7	4	6	4	7	4

Act. = Actual; Imp. = Improved

Session gaps (see Table VI) improved significantly, with minimum gaps increasing to 3 timeslots in October 2022, 6 in March 2023, and 15 in October 2023, indicating that 2-hour sessions are spread across 1–2 days. Reducing daily consecutive hours further contributed to more efficient timetables across all three semesters. Consequently, it provides a more balanced distribution of morning and afternoon timeslots across all groups' timetables (see Fig. 7).

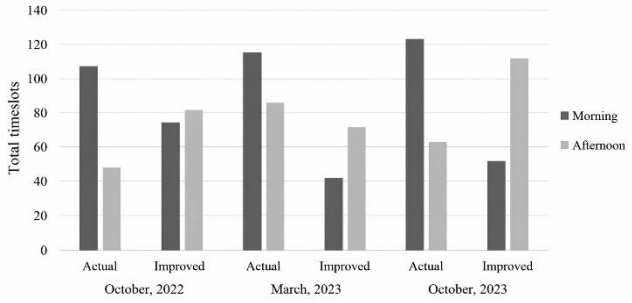


Fig. 7. Morning and afternoon timeslot distributions (all groups).

From the lecturer's timetable perspectives (Fig. 8), the distributions of morning and afternoon timeslots are more balanced in October 2022 and March 2023. In contrast, achieving this balance in October 2023 is difficult due to the higher complexity of multiple groups with more MAT/STA courses and shared-teaching constraints.

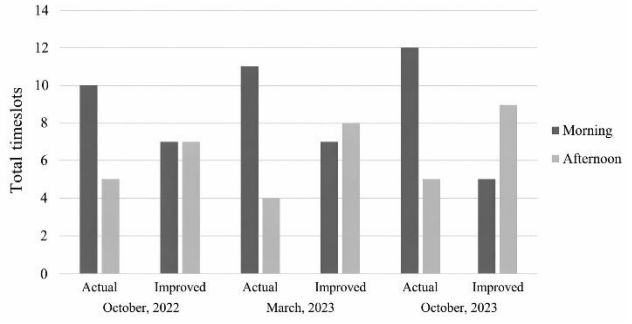


Fig. 8. Morning and afternoon timeslot distributions (all lecturers).

As a baseline comparison, 10 instances from ITC2002 and 12 from ITC2007 (Track 2) were evaluated. The results are compared with existing methods [2], [9], [17], which applied variants of evolutionary algorithms to the same benchmark datasets. The performance of benchmark datasets is measured by distance to feasibility (number of unallocated events) and the penalty cost of soft constraint violations. The soft constraints are included as follows:

- Students should not be scheduled to attend an event in the last timeslot of a day (that is, timeslots 9, 18, 27, 36, or 45).
- Students should not have to attend three (or more) events in consecutive timeslots occurring on the same day.
- Students should not be required to attend only one event on a particular day.

For the ITC2002 dataset, certain constraints are relaxed due to its slightly different characteristics, including session gaps, precedence events, and available timeslots. The proposed algorithm produces feasible solutions with no unallocated events for all instances (see Table VII). However, the resulting soft-constraint penalty costs are less competitive than those of previous works (see Table VIII).

TABLE VII. COMPARISON OF UNALLOCATED COURSES ON ITC2002 BENCHMARKS.

Instance	Proposed EA	[17]	[2]
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0

TABLE VIII. COMPARISON OF PENALTY COST ON ITC2002 BENCHMARKS

Instance	Proposed EA				[12]	[2]
	Initial Avg.	*Imp. Avg.*	*Initial Best*	*Imp. Best*		
1	599	444	566	407	52	77
2	585	443	544	405	20	250
3	568	534	540	512	78	100
4	831	674	783	611	74	130
5	922	806	842	704	71	326
6	886	767	837	681	6	359
7	1234	1114	1159	1049	6	574
8	680	545	628	490	15	383
9	630	507	586	472	32	256
10	589	499	516	460	58	73

Avg. = Average; Imp. = Improved

Meanwhile, constraints for ITC2007 (Track 2) are more complex than those for ITC2002, as events are allocated to specific time slots only. In addition, there is a precedence requirement that certain events should occur before certain others. Nevertheless, the session gaps are also relaxed during

testing of the ITC2007 (Track 2) instance. Consequently, only three instances (3, 7, and 8) can achieve zero unallocated events by the proposed algorithm (see Table IX).

TABLE IX. COMPARISON OF UNALLOCATED COURSES ON ITC2007 (TRACK 2) BENCHMARKS.

Instance	Proposed EA	[2]	[9]
1	18	0	0
2	37	0	0
3	**0**	**0**	**0**
4	2	0	0
5	34	0	0
6	24	0	0
7	**0**	**0**	**0**
8	**0**	**0**	**0**
9	34	0	0
10	26	0	0
11	5	0	0
12	10	0	0

Furthermore, Table X presents a comparison of soft-constraint violations (penalty costs) for ITC2007 (Track 2). The penalty costs for initialization and improvement (Swapping 1 and Swapping 2) are compared with those in previous studies. The improved results consistently outperformed the initial solutions, demonstrating that the proposed variation operator significantly improves solution quality. Even though the proposed approach did not always outperform existing methods such as EM-GD [17], HH-PSO [2], and IPGALS [9], the observed improvements from the initial stage to the improved stage confirm that the algorithm can adapt and refine the solution effectively. This demonstrates robustness in maintaining quality, even at larger problem sizes.

TABLE X. COMPARISON OF PENALTY COST ON ITC2007 (TRACK 2) BENCHMARKS.

Instance	Proposed EA				[2]	[9]	[9]
	Initial Avg.	*Imp. Avg.*	*Initial Best*	*Imp. Best*		*Avg.*	*Best*
1	974.86	735.9	822	**417**	70	437	**409**
2	1073.6	857.35	1011	813	50	397	381
3	2544.55	1084.1	2356	996	150	214	195
4	2393.75	1108.2	2110	992	228	242	211
5	845.5	455	792	424	4	0	0
6	1028.55	387.05	986	353	0	0	0
7	1778.7	348.85	1682	320	0	0	0
8	2179.25	329.95	2095	300	0	0	0
9	1011.55	863.55	927	**798**	**1794**	0	0
10	1067.8	795.9	992	750	160	654	476
11	2155.4	1150.85	1935	1104	230	149	135
12	2144.65	1115	1916	1023	34	169	153

Avg. = Average; Imp. = Improved

IV. CONCLUSION

In this study, an evolutionary algorithm with a dynamic local operator was used to allocate MAT/STA courses within the Mathematical Sciences Department at UiTM Sarawak Branch. The dynamic local operator was used to relocate and recompute the session gaps for 2+2 lecture hours of MAT/STA courses while satisfying a balanced distribution within a pre-scheduled timetable (limited available time slots).

Computational results are compared with the manual solutions for three semesters. It shows that the proposed algorithm can handle more complex scenarios (multiple groups with more MAT/STA courses) at the cost of a later end time. The proposed variation operator significantly improves solution quality, even when a good initial population has already been generated (compared to manual solutions). On average, the penalty value is reduced by up to 66.30%, and up to 15 timeslots improve the session gap. The proposed algorithm was further tested on ITC2002 and ITC2007 (Track 2) to serve as a baseline for existing methods. With certain constraint relaxations, the benchmark dataset results indicate that the proposed algorithm maintains solution quality even in larger, more complex settings.

ACKNOWLEDGMENT

This research was partially funded by the Universiti Malaysia Sarawak (UNIMAS).

REFERENCES

[1] S. Imran Hossain, M. A. H. Akhand, M. I. R. Shuvo, N. Siddique, and H. Adeli, "Optimization of University Course Scheduling Problem using Particle Swarm Optimization with Selective Search," Expert Syst. Appl., vol. 127, pp. 9–24, Aug. 2019, doi: 10.1016/j.eswa.2019.02.026.

[2] Z. Iqbal, R. Ilyas, H. Y. Chan, and N. Ahmed, "Effective Solution of University Course Timetabling using Particle Swarm Optimizer based Hyper Heuristic approach," Baghdad Science Journal, vol. 18, no. 4, pp. 1465–1475, 2021, doi: 10.21123/bsj.2021.18.4(Suppl.).1465.

[3] T. Thepphakorn, S. Sooncharoen, and P. Pongcharoen, "Particle Swarm Optimisation Variants and Its Hybridisation Ratios for Generating Cost-Effective Educational Course Timetables," SN Comput. Sci., vol. 2, no. 4, Jul. 2021, doi: 10.1007/s42979-021-00652-2.

[4] I. A. Abduljabbar and S. M. Abdullah, "An evolutionary algorithm for solving academic courses timetable scheduling problem," Baghdad Science Journal, vol. 19, no. 2, pp. 399–408, 2022, doi: 10.21123/BSJ.2022.19.2.0399.

[5] E. C. Perez, O. M. Rios, D. P. Bautista, S. S. Sanchez, and F. A. Acevedo, "A Genetic Algorithm Solution for Scheduling Problem," in CONIIN 2021 - 17th International Engineering Congress, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/CONIIN54356.2021.9634725.

[6] J. S. Tan, S. L. Goh, S. Sura, G. Kendall, and N. R. Sabar, "Hybrid particle swarm optimization with particle elimination for the high school timetabling problem," Evol. Intell., vol. 14, no. 4, pp. 1915–1930, Dec. 2021, doi: 10.1007/s12065-020-00473-x.

[7] C. Altıntaş and T. Yiğit, "Analysing the effects of classroom utilisation with a self-generating multimeme memetic algorithm for the exam timetabling problem," Advances in Artificial Intelligence Research (AAIR), vol. 1, no. 2, pp. 43–51, 2021, [Online]. Available: https://dergipark.org.tr/tr/pub/aair

[8] D. F. Dofadar, R. H. Khan, S. Hasan, T. A. Taj, A. Shakil, and M. Majumdar, "A Hybrid Evolutionary Approach to Solve University Course Allocation Problem," in Proceedings - 2021 International Conference on Artificial Intelligence and Blockchain Technology, AIBT 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 48–52. doi: 10.1109/AIBT53261.2021.00015.

[9] A. Rezaeipanah, S. S. Matoori, and G. Ahmadi, "A hybrid algorithm for the university course timetabling problem using the improved parallel genetic algorithm and local search," Applied Intelligence, vol. 51, no. 1, pp. 467–492, Jan. 2021, doi: 10.1007/s10489-020-01833-x.

[10] G. K. Edmund K. Burke, Search Methodologies. Boston, MA: Springer US, 2014. doi: 10.1007/978-1-4614-6940-7.

[11] A. Hamdan, S. San Nah, G. Say Leng, C. Kang Leng, and T. Wei King, "Recent Evolutionary Algorithm Variants for Combinatorial Optimization Problem," Applications of Modelling and Simulation, vol. 7, pp. 214–238, Dec. 2023, [Online]. Available: https://arqiipubl.com/ojs/index.php/AMS_Journal/article/view/515

[12] R. Lewis, "A survey of metaheuristic-based techniques for University Timetabling problems," OR Spectrum, vol. 30, no. 1, pp. 167–190, Jan. 2008, doi: 10.1007/s00291-007-0097-0.

[13] R. Lewis, "A time-dependent metaheuristic algorithm for post enrolment-based course timetabling," Ann. Oper. Res., vol. 194, no. 1, pp. 273–289, Apr. 2012, doi: 10.1007/s10479-010-0696-z.

[14] H. Cambazard, E. Hebrard, B. O'Sullivan, and A. Papadopoulos, "Local search and constraint programming for the post enrolment-based course timetabling problem," Ann. Oper. Res., vol. 194, no. 1, pp. 111–135, Apr. 2012, doi: 10.1007/s10479-010-0737-7.

[15] C. Nothegger, A. Mayer, A. Chwatal, and G. R. Raidl, "Solving the post enrolment course timetabling problem by ant colony optimization," Ann. Oper. Res., vol. 194, no. 1, pp. 325–339, Apr. 2012, doi: 10.1007/s10479-012-1078-5.

[16] G. Kobeaga, M. Merino, and J. A. Lozano, "An efficient evolutionary algorithm for the orienteering problem," *Comput. Oper. Res.*, vol. 90, pp. 42–59, Feb. 2018, doi: 10.1016/j.cor.2017.09.003.

[17] S. Abdullah, H. Turabieh, B. McCollum, and P. McMullan, "A hybrid metaheuristic approach to the university course timetabling problem," Journal of Heuristics, vol. 18, no. 1, pp. 1–23, Feb. 2012, doi: 10.1007/s10732-010-9154-y.

Multi-Class Stress Detection Using Electrodermal Activity: Evaluation of A Hybrid Model Using UBFC-Phys Dataset

Kawther Alsayed, Hamza Ghandorh, Wael M.S. Yafooz
Department of Computer Science-College of Computer Science and Engineering, Taibah University, Medinah, Saudi Arabia

***Abstract*—Stress is a psychological and physiological response to internal or external pressures or challenges that exceed an individual's ability to cope. In response to these conditions, the human body produces physiological signals that reflect its internal states, often without conscious awareness. Among these signals, electrodermal activity (EDA) is known for its sensitivity to changes in stress levels. Traditional methods for assessing stress, such as questionnaires and self-reports, remain widely used, but they are often influenced by subjectivity and recall bias. This has led to increased interest in objective, data-driven approaches. This research proposes a hybrid stress detection model. The EDA signals were used from the UBFC-Phys dataset, which comprises data from 56 participants under three conditions: rest, moderate stress, and high stress. The signals were preprocessed using filtering, smoothing, and subject-level normalization to reduce inter-individual variability. A set of features was extracted, such as tonic component features. Multiple experiments were conducted on baseline, deep learning, and hybrid models not only for stress detection classification purposes but also for performance evaluation purposes. The proposed model achieved an accuracy of 92.16% in the multi-class classification task. The findings of this research contribute to mental health and well-being by providing an accurate model for stress detection that can be adapted to healthcare, education, and workplace settings, ensuring a healthier and more sustainable future.**

***Keywords*—Stress detection; physiological signal; Electrodermal Activity (EDA); machine learning; ensemble; deep learning; hybrid model**

I. Introduction

Stress negatively affects both physical and mental health. It is causative of conditions such as cardiovascular diseases, weakened immunity, anxiety, depression, and sleep disorders [1], [2]. Its effects are not limited to the individual level but extend to productivity and quality of life in general, such as work environments and educational institutions [3], [4]. This highlights the importance of identifying stress early and monitoring it over time to support overall well-being [5]. When a person experiences psychological stress, the autonomic nervous system, particularly the sympathetic system, is triggered. As a result, hormones such as cortisol and adrenaline are released to prepare the body for a fight-or-flight response [6]. These hormones increase blood pressure, heart rate, and breathing rate, and lead to noticeable changes in physiological signals [6]. After the stress ends, the parasympathetic nervous system helps the body recover and return to a normal state.

Recently, more attention has been given to using physiological signals to detect stress directly and continuously. One of the most common signals is electrodermal activity (EDA), a physiological signal that captures changes in skin conductance associated with sweat gland activation by the sympathetic nervous system [7]. Current progress in machine learning (ML) and deep learning (DL) makes it possible to automatically analyze physiological signals and classify stress levels. Various ML models have been applied to learn patterns from extracted features. DL approaches, e.g., Convolutional Neural Networks (CNN) and Transformers, have also been used to capture more complex relationships within the data.

Despite recent progress in stress detection, working with the UBFC-Phys dataset [8] still presents a challenge. This dataset requires careful handling due to the nature of the physiological signals. There are clear differences between individuals, as each person has their own baseline and responds differently to stress. In addition, some patterns, especially between moderate and high stress, can be very similar, which makes classification more challenging. Thus, it is important to use methods that take these differences into account and provide consistent results across all states. The main contributions of this work are:

Addressing inter-individual variability by applying subject-level normalization, which reduces differences in physiological baselines and improves the consistency of stress classification.

Proposed a hybrid model for binary and multiclass stress classification and achieved high performance on the UBFC-Phys dataset.

The remainder of this study is organized as follows: Section II presents a review of the relevant literature. Section III details the materials and methods, including the dataset description, preprocessing procedures, feature extraction process, data partitioning strategy, proposed model architecture, and evaluation metrics. Section IV presents the experiments and results, followed by a comparison with previous work. Section V discusses the findings. Section VI concludes the study.

II. Literature Review

Recent studies have shown increasing interest in stress detection using physiological signals, as these signals provide more objective indicators. Many publicly available datasets have been widely used in stress detection research to develop, train, and evaluate various ML and DL models using physiological signals. These datasets provide important

benchmarks for comparing stress detection models in different environments and with different signals. The UBFC-Phys dataset was used in this research. Other datasets include SWELL-KW [9], which is designed to evaluate workplace stress under time-pressure and email interruption conditions, and WESAD [10], which provides a multimodal dataset that involves physiological signals recorded from chest- and wrist-worn devices during different conditions. Other datasets that contribute to stress detection research comprise SRAD [11], Louisiana [12], TILES-2018 [13], and MAUS [14], which provide data related to driving stress, hospital work environments, workplace stress, and mental workload.

Physiological signals are used for stress detection with the support of wearable devices and AI models. Different ML approaches have been applied in this area. Sabour et al. presented the UBFC-Phys dataset and used EDA and BVP signals to study social stress, showing their usefulness in this field [8]. Vos et al. developed an ensemble model using data from multiple sources and showed good generalization for stress prediction [15]. Hasanpoor et al. used PPG signals from the UBFC-Phys dataset with a CNN-MLP model and reported an accuracy of 82%, showing that a single signal can be used for stress detection [16].

Zhang et al. used a DL model based on attention to combine BVP and EDA signals and capture the relationship between them [17]. Arsalan et al. [18] applied several supervised learning models with EEG, GSR/EDA, and PPG signals, and reported that the MLP model achieved the highest accuracy. Al Hasan et al. [19] used PPG, ECG, and GSR from the MAUS dataset and reported good performance using the Random Forest (RF) model after hyperparameter tuning. Mathur et al. [20] investigated stress detection among nurses. In their experiment, they used wearable sensor data and achieved high performance with Decision Tree classifiers.

WESAD and other datasets have been used to compare ML and DL models for stress detection. Kar et al. used a GRU-RNN model with EDA, accelerometer, and temperature signals and achieved high performance in binary classification [21]. Gil-Martin et al. applied CNN models using data from wearable sensors and showed that DL can learn useful patterns from physiological signals [22]. Liu et al. used a Transformer-LSTM model with PPG signals and reported strong results in stress classification [23]. Other work explored CNN-LSTM models and multimodal signals to improve performance, as combining different signals helps capture more information about the body's response to stress [24]. These studies show that DL models can handle complex patterns, especially when enough data is available.

Many physiological signals were used for stress detection, such as electrodermal activity (EDA), blood volume pulse (BVP), electrocardiogram (ECG), photoplethysmography (PPG), and heart rate variability (HRV) [25]. Electrodermal activity (EDA), on the other hand, is a practical option because it is directly related to sympathetic nervous system activity and can be collected using wearable devices [7]. Despite these improvements, distinguishing between different stress levels is still challenging, especially when the patterns are very similar. Based on this, the current study focuses on using Electrodermal activity (EDA) signals from the UBFC-Phys dataset and proposes a hybrid model for both binary and multiclass stress classification.

III. Methodology

A. Dataset

Data were obtained from a publicly accessible dataset named UBFC-Phys [8]. The dataset contains physiological signals collected using the E4 wristband EDA and BVP together with video recordings from 56 participants during three tasks performed under different stress conditions. The experiment included three tasks representing different stress levels: a resting task (T1) used as the baseline condition, a speech task (T2), and an arithmetic task (T3). Each task included 56 recordings, resulting in a total of 168 signals. This work focused on the EDA signals. For each subject, the signals were recorded separately for each task, which made it easier to distinguish between rest and stress states.

Fig. 1 shows that EDA signals increased with higher stress levels across the tasks. The signal was stable during rest (T1), slightly higher during the speaking task (T2), and highest during the arithmetic task (T3), showing that EDA can reflect changes in stress levels.

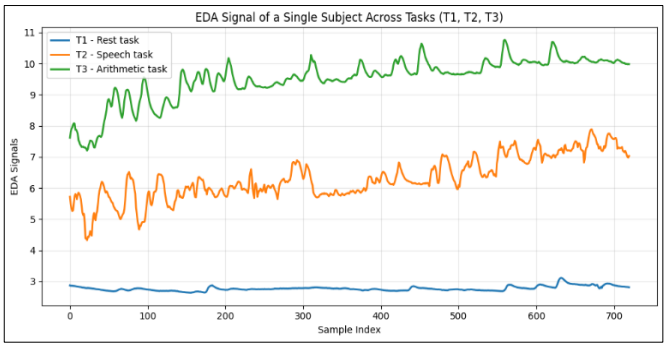


Fig. 1. EDA signal of Subject 13 across the three tasks: rest (T1), speech (T2), and arithmetic (T3). The signal amplitude increases from the resting condition to the more demanding tasks, reflecting higher physiological arousal under stress.

B. Data Preprocessing

The EDA signals were recorded at a sampling rate of 8 Hz. Several preprocessing steps were applied to reduce noise, enhance the quality of the signal, and prepare the data. First, the initial two seconds of each signal were removed, as this part often contained unstable values due to sensor contact at the beginning of the recording.

A fourth-order Butterworth low-pass filter was applied to attenuate high-frequency artifacts with a 1 Hz cutoff frequency, while keeping the main signal pattern unchanged. A Savitzky-Golay filter was used to smooth and reduce small fluctuations of the signal without affecting the overall shape.

The dataset showed variability between participants in their physiological responses. To address this, Z-score normalization was applied for each subject to reduce these differences and allow the model to focus on relative changes in the EDA signal rather than absolute values. The mean and standard deviation (SD) were calculated from the resting condition (T1) only and

subsequently used to normalize all recordings for each participant across different tasks.

$$x_norm = \frac{(x - \mu_baseline)}{\sigma_{baseline} + \varepsilon} \quad (1)$$

where x was the EDA value, μ_baseline and σ_baseline were the mean and standard deviation of the resting signal (T1), and ε is a small constant.

C. Feature Extraction

For each EDA signal, 23 features were extracted to represent different signal characteristics. The extracted features were directly used for model classification. The tonic and phasic components of each EDA signal were extracted using NeuroKit2 [26]. Six features were extracted from the tonic component to describe the baseline level, variability, and overall trend of the signal. Similarly, six features were derived from the phasic component to capture rapid changes related to sympathetic activity.

Additionally, seven statistical features were computed directly from the preprocessed signal to represent its overall distribution and variability. Two Hjorth [27] parameters were also included to describe the signal dynamics and complexity. The zero-crossing rate of the first derivative was used to reflect local variations in the signal, and low-frequency spectral power was calculated to capture the dominant physiological patterns. A summary of extracted features is shown Table I.

TABLE I. FEATURES DERIVED FROM EDA SIGNALS [8]

Feature	Description
scl_mean	Mean of skin conductance level (SCL)
scl_std	Tonic component standard deviation
scl_range	Range (max − min) of the tonic component
scl_slope	Linear trend (slope) of the tonic component over time
scl_median	Median value of the tonic component
scl_cv	Coefficient of variation of the tonic component
scr_mean	Mean absolute value of the phasic (SCR) component
scr_std	Standard deviation of the phasic component
scr_max	Maximum value of the phasic component
scr_auc	Area under the positive phasic component curve
scr_energy	Total energy of the phasic component
scr_count	Number of detected SCR peaks
raw_mean	Mean of the raw EDA signal
raw_std	Standard deviation of the raw EDA signal
raw_var	Variance of the raw EDA signal
raw_range	Range (max - min) of the raw EDA signal
raw_skew	Skewness of the raw EDA signal distribution
raw_kurt	Kurtosis of the raw EDA signal distribution
raw_rms	Root mean square of the raw EDA signal
hjorth_mobility	Rate of change of the signal relative to its variance
hjorth_complexity	Rate of change of the signal's frequency content
zcr	Rate of sign changes in the first derivative of the signal
lf_power	Power spectral density in the low-frequency band (0-1 Hz)

D. Data Splitting

The dataset was partitioned into training and testing sets using a 70:30 ratio to enable reliable model evaluation. A subject-independent split was applied using GroupShuffleSplit with a fixed random state (42), where all recordings from each participant were assigned to either the training or testing set to avoid data leakage. Within the training set, hyperparameter tuning was conducted using five-fold cross-validation. GroupKFold was used for binary classification, and StratifiedGroupKFold was applied for multiclass tasks. GridSearchCV was used to select the best hyperparameters, and the final model was evaluated on the unseen data (test set).

E. Experimental Setup

The experimental setup in this work was organized into three experiments.

1) Binary classification using ML models: The first experiment evaluated baseline ML models for binary stress classification using EDA-based features. T1 and T2 from the UBFC-Phys dataset were used, where T1 represented the non-stress state, and T2 represented the stress state. The experiment was conducted under different preprocessing settings, without smoothing and normalization and with smoothing and normalization, to examine their effect on classification performance. Fig. 2 shows an overview of the first experiment for binary stress classification. The configurations used in the experiment are summarized in Table II.

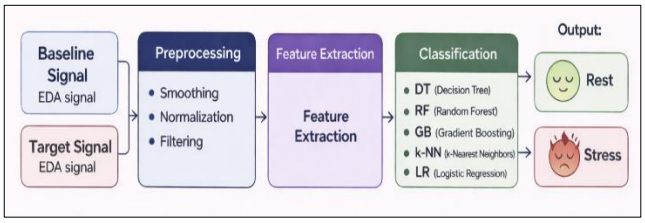


Fig. 2. Overview of the binary Stress Detection Pipeline. It starts with EDA signals (Baseline and Target), then preprocesses them. After that, features are extracted and fed into ML models to classify the stress.

TABLE II. HYPERPARAMETER SETTINGS FOR BASELINE ML MODELS IN BINARY CLASSIFICATION

Model	HyperParameter
GB	n_estimators = 100, 150, 200; learning_rate = 0.01, 0.05, 0.1; max_depth = 2, 3
RF	n_estimators = 200, 300, 500; max_depth = 5, 10, 15, 20, None; min_samples_split = 2, 5; min_samples_leaf = 1, 2
K-NN	n_neighbors = 3, 5, 7, 9, 11; weights = uniform, distance; metric = euclidean, manhattan, minkowski
LR	C = 0.001, 0.01, 0.1, 0.5, 1, 5, 10; max_iter = 2000
DT	max_depth = 3, 5, 10, 15, None; min_samples_split = 2, 5, 10; criterion = gini, entropy

2) Multiclass classification using ML and DL models: The second experiment evaluated multi-class stress classification using both ML and DL models. Three tasks from the dataset were used, including T1 for rest, T2 for moderate stress, and T3 for high stress. The same preprocessing steps, feature extraction process, and data splitting strategy used in the first experiment

were also applied (Fig. 3). The configurations used in the experiment are summarized in Table III.

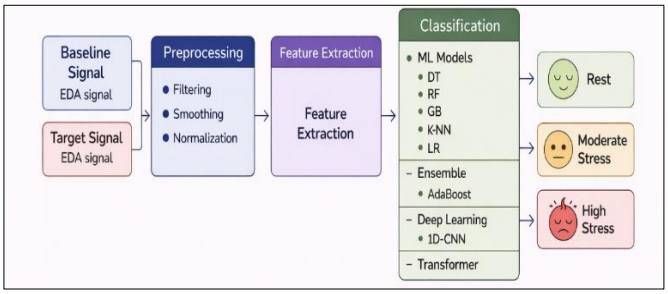


Fig. 3. Overview of the Multi-Level Stress Detection Pipeline: From raw EDA signals, the data preprocessing, and then feature extraction. Different models then use the extracted features to predict three stress levels.

TABLE III. HYPERPARAMETER SETTINGS FOR MODELS IN MULTICLASS CLASSIFICATION

Model	HyperParameter
GB	n_estimators = 100, 150, 200; learning_rate = 0.01, 0.05, 0.1; max_depth = 2, 3
RF	n_estimators = 200, 300; max_depth = 10, 20, None
K-NN	n_neighbors = 3, 5, 7, 9; weights = uniform, distance
LR	C = 0.01, 0.1, 1, 10; solver = lbfgs
DT	max_depth = 3, 5, 10, None; criterion = gini, entropy
AdaBoost	base_estimator = Decision Tree (max_depth = 2), SAMME; n_estimators = 50, 100, 200; learning_rate = 0.01, 0.1, 1.0
1D-CNN	Filters = 64 / 128; FC Hidden = 64; Dropout = 0.4; Learning Rate = 0.0005; Epochs = 150; Batch Size = 16
Feature Transformer	d_model = 64; Attention Heads = 8; Encoder Layers = 2; FC Hidden = 64; Dropout = 0.3; Learning Rate = 0.0005; Epochs = 200; Batch Size = 16; Noise Std = 0.02

3) Proposed hybrid model for multiclass classification: The proposed hybrid model combined five ML classifiers, Gradient Boosting (GB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), using a weighted soft voting strategy. GB was given the highest weight due to its strong capability for sequential error correction. RF provided robustness. KNN contributed by capturing local data structures, LR introduced a linear decision perspective, and DT added model diversity with low complexity. Fig. 4 shows an overview of the proposed model.

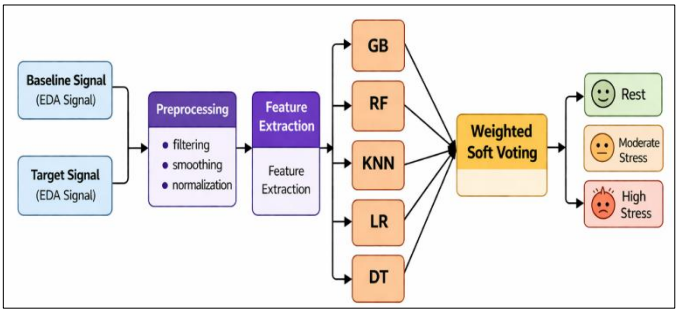


Fig. 4. Overview of the hybrid Stress Detection Pipeline for Multiclass: It starts with EDA signals, followed by preprocessing, feature extraction, and the hybrid model predicts three stress levels: Rest, Moderate Stress, and High Stress.

The configurations used in the experiment are summarized in Table IV.

TABLE IV. THE PROPOSED HYBRID MODEL CONFIGURATION

Model	HyperParameter	Weight
GB	n_estimators = 150, learning_rate = 0.01, max_depth = 2	3
RF	n_estimators = 300, max_depth = 10, min_samples_split = 5	2
K-NN	n_neighbors = 7, weights = distance	2
LR	C = 10, solver = lbfgs	1
DT	criterion = gini, max_depth = 3	1

F. Performance Evaluation

Several evaluation metrics were employed to assess various aspects of model performance. The models were evaluated on the test set using widely adopted performance measures, including accuracy, recall, precision, F1-score, ROC curves, and the confusion matrix. The performance metrics utilized in this work are summarized in Table V.

TABLE V. EVALUATION METRICS UTILIZED FOR PERFORMANCE ASSESSMENT

Metric	Description
PR	$PR = \frac{TP}{TP + FP} \times 100$
RE	$RE = \frac{TP}{TP + FN} \times 100$
AC	$AC = \frac{\mathrm{TP} + \mathrm{TN}}{Total\ Samples} \times 100$
F1	$F1 - score = 2 \times \frac{PR\ x\ RE}{PR + RE}$
CM	Provides class-wise information on correct classifications and misclassification patterns.
ROC	Shows how the model performance changes across different decision thresholds

IV. RESULTS

In the first experiment, baseline ML models were evaluated for binary stress classification without applying smoothing and normalization. Table VI shows the performance results of the models using EDA-based features for classifying between stress and non-stress states.

TABLE VI. THE RESULTS OF THE BASELINE EXPERIMENT (WITHOUT SMOOTHING AND NORMALIZATION)

Model	AC	PR	RE	F1
KNN	**91.18**	**85.00**	**100**	**91.89**
GB	88.24	80.95	100	89.47
LR	85.29	83.33	88.24	85.71
RF	82.35	76.19	94.12	84.21
DT	82.35	82.35	82.35	82.35

Fig. 5 and Fig. 6 show the confusion matrices for the selected models in the first experiment. Both GB and KNN achieved high recall for stress, with some false positives where rest samples were misclassified as stress. KNN showed slightly fewer misclassifications compared to GB.

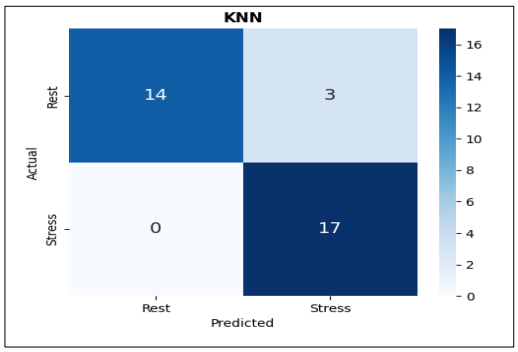


Fig. 5. CM of binary classification (no smooth/no norm) for KNN model.

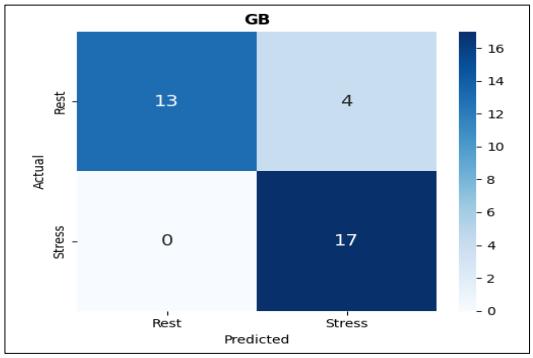


Fig. 6. CM of binary classification (no smooth/no norm) for GB model.

The ROC curves in Fig. 7 and 8 below show that KNN achieved the highest AUC (0.976), which indicated the best separation between the two classes. GB also had a high AUC (0.934), showing good performance.

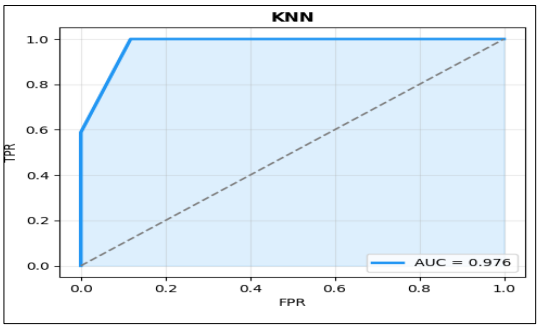


Fig. 7. ROC curve of binary classification (no smooth/no norm) for KNN model.

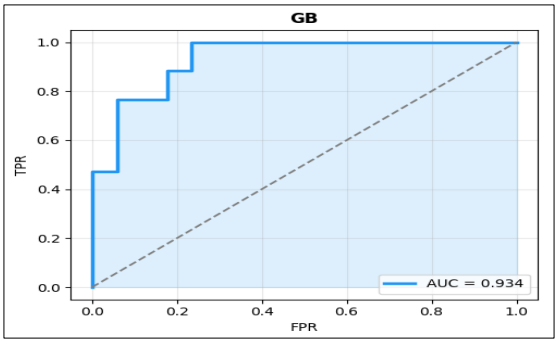


Fig. 8. ROC curve of binary classification (no smooth/no norm) for GB model.

Table VII presents the results after applying smoothing and normalization. The preprocessing steps improved the performance of most models compared to the previous setting. DT achieved perfect results. These results show that the preprocessing pipeline helped the models better distinguish between the classes and improved generalization across different subjects. The high performance may also be related to the relatively small dataset size and the simpler nature of the binary classification task.

TABLE VII. THE RESULTS OF THE BINARY CLASSIFICATION (WITH SMOOTHING AND NORMALIZATION)

Model	AC	PR	RE	F1
DT	**100**	**100**	**100**	**100**
RF	97.06	100	94.12	96.97
GB	97.06	100	94.12	96.97
KNN	91.18	85.00	100	91.89
LR	88.24	88.24	88.24	88.24

Fig. 9 and 10 show the confusion matrices for DT and RF in binary classification using preprocessed EDA signals. DT correctly classified all samples without errors, while RF also showed strong performance with only one misclassification.

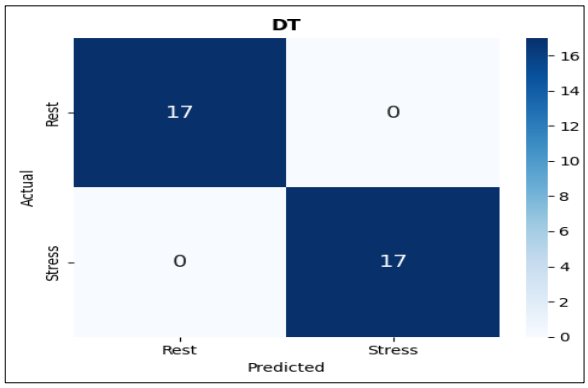


Fig. 9. CM of binary classification (with smooth and norm) for DT model.

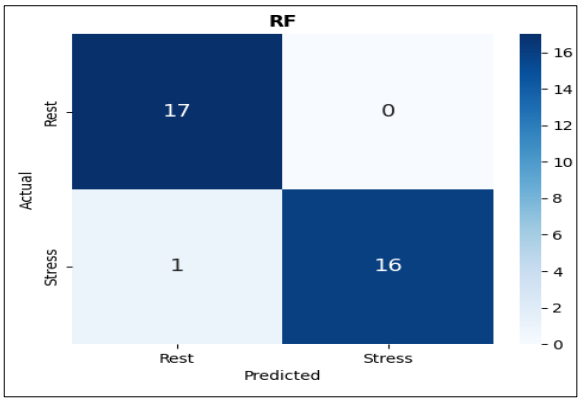


Fig. 10. CM of binary classification (with smooth and norm) for RF model.

Fig. 11 and 12 show the ROC curves for both DT and RF, which achieved perfect AUC values of 1.000, indicating a clear separation between the rest and stress classes.

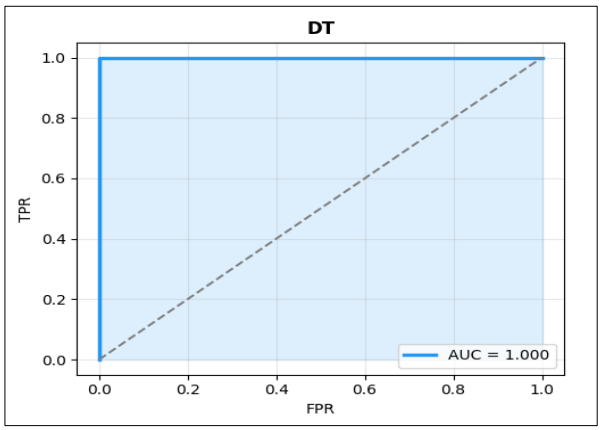


Fig. 11. ROC curve of binary classification (with smooth and norm) for DT model.

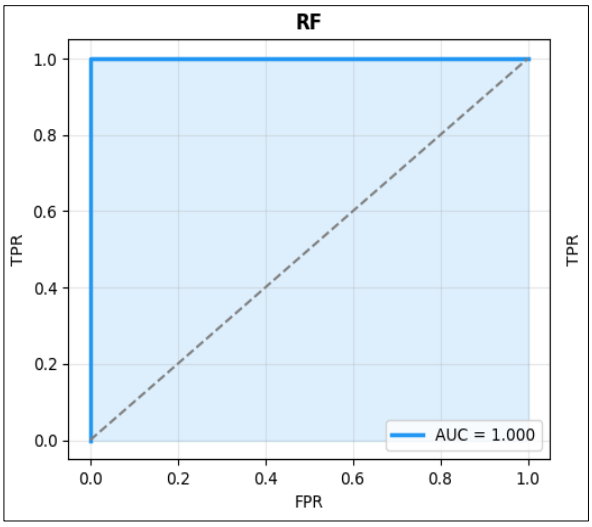


Fig. 12. ROC curve of binary classification (with smooth and norm) for RF model.

The second experiment evaluated multi-class stress classification using both ML and DL models. The three-level stress classification results are presented in Table VIII. The best performance was achieved by GB, while comparable results were obtained by RF, KNN, AdaBoost, and DT. Lower performance was observed for LR and Transformer, whereas the weakest results were recorded by 1D-CNN.

TABLE VIII. RESULTS OF THE MULTICLASS CLASSIFICATION

Model	AC	PR	RE	F1
GB	**90.20**	**90.28**	**90.20**	**90.19**
RF	88.24	88.24	88.24	88.24
KNN	88.24	88.24	88.24	88.24
AdaBoost	88.24	88.24	88.24	88.24
DT	88.24	88.24	88.24	88.24
LR	86.27	86.73	86.27	86.36
Transformer	82.35	82.41	82.35	82.34
1D-CNN	52.94	47.15	52.94	44.23

The relatively low performance of the 1D-CNN might be due to the small size of the dataset and the subject-independent split, whereby the number of samples available for deep feature learning was reduced. The three classes were well balanced, but the low-frequency EDA and the subtle differences between moderate and high stress could have affected the discriminative representation that the 1D-CNN was able to learn. Therefore, the poor performance of the lower 1D-CNN can be attributed to the model architecture used and less training data.

The confusion matrices for the selected models are shown in Fig. 13 and 14 below show the confusion matrices for the selected models. The GB model correctly classified all the rest of the samples and achieved strong performance across stress levels. The Feature Transformer also correctly identified all the remaining samples but showed more confusion between stress levels compared to GB.

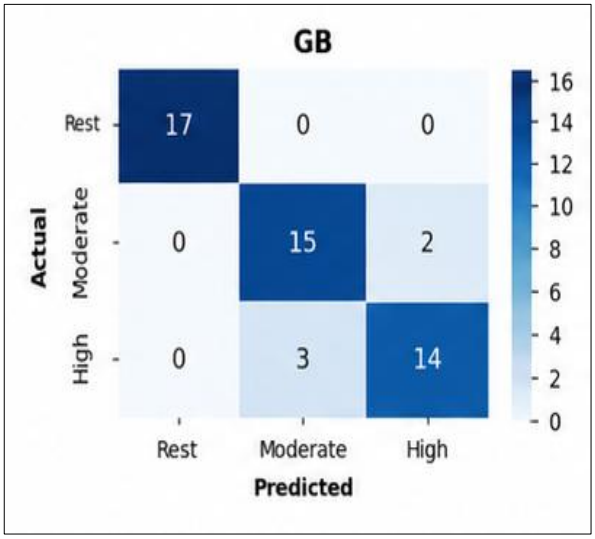


Fig. 13. CM of multiclass classification for the GB model.

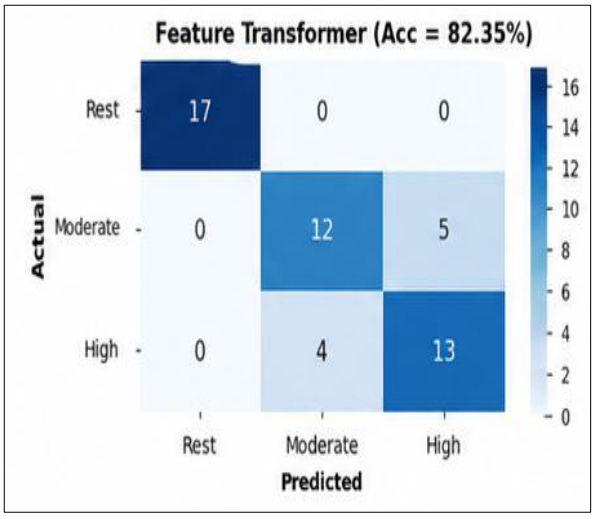


Fig. 14. CM of multiclass classification for Transformer.

Fig. 15 and 16 below show the ROC curves for the selected models. Both models achieved strong performance with clear separation for the rest of the classes.

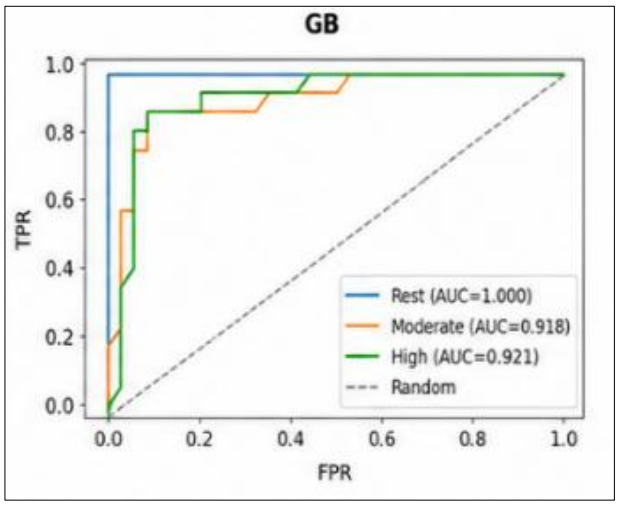


Fig. 15. ROC curve of multiclass classification for GB model.

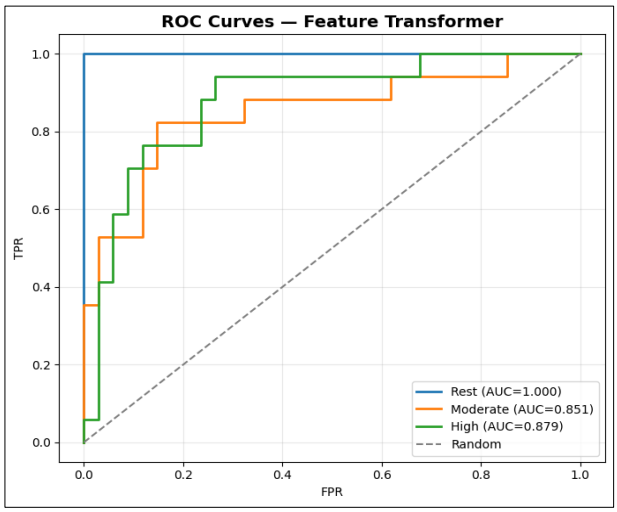


Fig. 16. ROC curve of multiclass classification for Transformer.

The final experiment evaluated the proposed hybrid model. Table IX presents the performance results of the proposed model. The proposed hybrid model outperformed all other evaluated models in terms of overall performance.

TABLE IX. THE RESULTS OF PROPOSED MODEL

Model	AC	PR	RE	F1
Proposed Hybrid Model	92.16	92.16	92.16	92.16

The confusion matrix in Fig. 17 shows that all the rest of the samples were correctly classified by the proposed hybrid model. Most moderate and high stress samples were also classified correctly, with only a few errors between them. To analyze the contribution of extracted EDA features, the proposed hybrid model was implemented with feature group ablation analysis. One feature group was eliminated in each experiment, and the other features were kept. The results are presented in Table X.

The full 23-feature set achieved the highest accuracy of 92.16%. Removing the phasic features resulted in the greatest drop in performance, indicating their importance in capturing stress-related sympathetic responses. Removing tonic and dynamic/spectral features also shows a reduction, further indicating that different groups of EDA features provide complementary information for multi-class stress classification.

TABLE X. FEATURE-GROUP ABLATION ANALYSIS OF THE PROPOSED HYBRID MODEL

Feature setting	Features	AC
Full feature set	**23**	**92.16**
Without tonic features	**17**	**88.24**
Without phasic features	**17**	**86.27**
Without statistical features	**16**	**90.20**
Without dynamic/spectral features	**19**	**88.24**

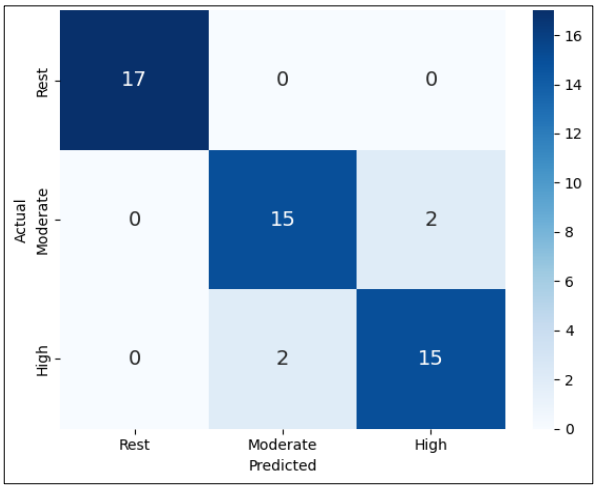


Fig. 17. CM of multiclass classification for the proposed hybrid model.

The ROC curves in Fig. 18 for the proposed hybrid model show a perfect AUC of 1.000 for the rest class. The high and moderate stress classes recorded AUCs of 0.946 and 0.943, respectively, confirming strong discrimination ability across all stress levels.

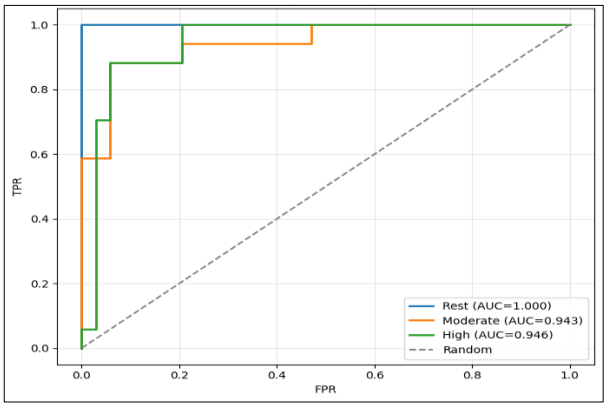


Fig. 18. ROC curve of multiclass classification for proposed hybrid model.

The result of the proposed model was compared with previous stress detection studies that used EDA signals on the UBFC-Phys dataset. Table XI compares the proposed model with EDA-based baseline models evaluated on the UBFC-Phys dataset.

TABLE XI. COMPARISON OF THE PROPOSED MODEL WITH EXISTING MULTICLASS STRESS DETECTION MODELS ON THE UBFC-PHYS DATASET

Reference	Method	Signal Used	AC
Zhang et al. [17]	DL (Attention)	EDA	48.1%
Sabour et al. [8]	SVM	EDA	63.60%
Valerio and Mahmoud [28]	SVM	EDA	69.12%
Proposed Model	**Weighted soft-voting**	**EDA**	**92.16%**

The comparison shows that the proposed hybrid model achieved the highest accuracy among the EDA-based models evaluated on the UBFC-Phys dataset. The performance of the proposed model was compared with the published SVM baseline and individual EDA-based ML/DL baselines for the multi-class stress classification task and outperformed both. This means that the proposed model is capable of capturing stress-related patterns without using any other signals except EDA signals.

V. DISCUSSION

The results of this work provide useful insight for computer science in ML, signal processing, and affective computing. The proposed model shows that a single physiological signal, EDA, could achieve high AC in stress classification when it was properly processed using steps such as filtering, smoothing, subject-level normalization, and feature extraction. This means that it was possible to build an effective stress detection system without relying on multiple sensing methods. EDA signals provided meaningful information for distinguishing between different stress levels. Separating closely related stress states could be challenging, especially between the speech task (T2) and the arithmetic task (T3), as they show similar patterns. In contrast, the binary task, which uses T1 and T2, was easier to distinguish, leading to higher performance. The proposed model achieved a high multiclass AC of 92.16%, showing that EDA could be effectively used as a standalone signal for stress detection. The improvement in performance highlights the importance of model design. The hybrid model combines multiple models, which helped capture different patterns in the data. This results in more stable and reliable predictions, especially when working with complex physiological signals. In addition, preprocessing steps such as filtering, smoothing, and subject-level normalization contribute to improving data quality and reducing differences between individuals. The subject-level normalization used in this work helps address one of the main challenges encountered, which is the variation between individuals. It was observed that different subjects have different physiological baselines, such as how their bodies responded to stress or how their skin reacted, which can affect model performance. By normalizing each subject's data based on their own rest state, the model becomes more adapted to each individual and less influenced by these differences. This makes the approach more reliable and more suitable for real-world applications, especially in personalized health monitoring systems that need to adapt to different users. This work highlights that carefully designed classical ML approaches could outperform more complex DL models, especially when working with relatively small datasets where DL models lack sufficient training samples to learn generalizable representations. This is an important insight, as it suggests that model simplicity and proper data handling could be more effective than increasing model complexity.

VI. CONCLUSION

Stress affects an individual's physical and psychological well-being, impacting their quality of life. Long-term exposure to stress has been related to a wide range of health problems. As individuals are increasingly exposed to elevated levels of stress in both personal and professional contexts. The need for an effective stress-monitoring model is more critical than ever. Early detection of stress enables timely intervention, reduces the risk of long-term health complications, and supports overall well-being. However, the traditional approaches for stress assessment, such as self-reported questionnaires, are often subjective and may not accurately reflect an individual's true physiological state. This limitation drives the growing interest in developing objective, data-driven approaches for stress detection using physiological signals. This research work proposed a hybrid model that detects stress using EDA signals from the UBFC-Phys dataset. The EDA signals were preprocessed using low-pass filtering and smoothing, followed by subject-level normalization to reduce inter-individual variability. A total of 23 features were extracted, and a hybrid weighted soft-voting ensemble combining GB, RF, KNN, LR, and DT was used to classify stress into three levels. The proposed model achieved an AC of 92.16% in the multiclass classification task, with balanced PR, recall, and F1 values of 92.16%. In the binary classification task, a perfect AC of 100% was achieved. This approach provided a practical and efficient solution for real-world stress monitoring and continuous health-monitoring applications. Future work may concentrate on enhancing and extending the proposed approach by assessing the pipeline on larger and more diverse datasets, and applying cross-dataset validation to offer stronger evidence of its generalizability. Incorporating additional physiological signals such as HR and BVP alongside EDA, within a multimodal fusion framework, may further enhance classification accuracy, particularly for distinguishing between closely related stress states. Investigating advanced DL architectures with transfer learning strategies may also help address the limitations observed with DL models in this study.

FUNDING

This scientific research study was derived from a research grant funded by Taibah University, Madinah, Kingdom of Saudi Arabia, with the grant number (1447-06).

DATA AVAILABILITY

The data presented in this work are openly available in the UBFC-Phys repository and can be accessed via [https://ieee-dataport.org/open-access/ubfc-phys-2].

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Taibah University and the contributors of the UBFC-Phys dataset.

REFERENCES

[1] Harvard Health Publishing, "Identifying and relieving stress," Harvard Health. Accessed: Nov. 16, 2025. [Online]. Available: https://www.health.harvard.edu/topics/stress

[2] American Psychological Association, "Stress effects on the body," APA. Accessed: Nov. 16, 2025. [Online]. Available: https://www.apa.org/topics/stress/body

[3] H. M. Morales-Fajardo et al., "Towards a Non-Contact Method for Identifying Stress Using Remote Photoplethysmography in Academic Environments," Sensors, vol. 22, no. 10, p. 3780, May 2022, doi: 10.3390/s22103780.

[4] J. C. Fruehwirth, M. E. Mazzolenis, M. A. Pepper, and K. M. Perreira, "Perceived stress, mental health symptoms, and deleterious behaviors during the transition to college," PLoS ONE, vol. 18, no. 6, p. e0287735, Jun. 2023, doi: 10.1371/journal.pone.0287735.

[5] H. Kerbage, O. Bazzi, W. El Hage, E. Corruble, and D. Purper-Ouakil, "Early interventions to prevent post-traumatic stress disorder in youth after exposure to a potentially traumatic event: A scoping review," in Healthcare, MDPI, 2022, p. 818. Accessed: Nov. 19, 2025. [Online]. Available: https://www.mdpi.com/2227-9032/10/5/818

[6] L. K. McCorry, "Physiology of the Autonomic Nervous System," Am J Pharm Educ, vol. 71, no. 4, p. 78, Sep. 2007, doi: 10.5688/aj710478.

[7] F. Rodrigues and H. Correia, "Semi-supervised and ensemble learning to predict work-related stress," J Intell Inf Syst, vol. 62, no. 1, pp. 77–90, Feb. 2024, doi: 10.1007/s10844-023-00806-z.

[8] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappé, and F. Yang, "UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress," IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 622–636, Jan. 2023, doi: 10.1109/TAFFC.2021.3056960.

[9] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The SWELL Knowledge Work Dataset for Stress and User Modeling Research," in Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul Turkey: ACM, Nov. 2014, pp. 291–298. doi: 10.1145/2663204.2663257.

[10] S. Chakraborty et al., "Deep Learning Based Stress Assessment Using PPG Signals from WESAD Dataset," in 2025 23rd IEEE Interregional NEWCAS Conference (NEWCAS), Jun. 2025, pp. 296–300. doi: 10.1109/NewCAS64648.2025.11106968.

[11] J. A. Healey and R. W. Picard, "Stress Recognition in Automobile Drivers." physionet.org, 2008. doi: 10.13026/C2SG6B.

[12] S. Hosseini et al., "A multimodal sensor dataset for continuous stress detection of nurses in a hospital," Sci Data, vol. 9, no. 1, p. 255, Jun. 2022, doi: 10.1038/s41597-022-01361-y.

[13] K. Mundnich et al., "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers," Sci Data, vol. 7, no. 1, p. 354, Oct. 2020, doi: 10.1038/s41597-020-00655-3.

[14] W.-K. Beh, Y.-H. Wu, An-Yeu, and Wu, "MAUS: A Dataset for Mental Workload Assessmenton N-back Task Using Wearable Sensor," Nov. 03, 2021, arXiv: arXiv:2111.02561. doi: 10.48550/arXiv.2111.02561.

[15] G. Vos, K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi, "Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices," Journal of Biomedical Informatics, vol. 148, p. 104556, Dec. 2023, doi: 10.1016/j.jbi.2023.104556.

[16] Y. Hasanpoor, K. Motaman, B. Tarvirdizadeh, K. Alipour, and M. Ghamari, "Stress detection using ppg signal and combined deep cnn-mlp network," in 2022 29th National and 7th International Iranian Conference on Biomedical Engineering (ICBME), IEEE, 2022, pp. 223–228. Accessed: Nov. 19, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10052957/

[17] X. Zhang et al., "Dynamic Alignment and Fusion of Multimodal Physiological Patterns for Stress Recognition," IEEE Transactions on Affective Computing, vol. 15, no. 2, pp. 685–696, Apr. 2024, doi: 10.1109/TAFFC.2023.3290177.

[18] A. Arsalan, M. Majid, S. M. Anwar, and U. Bagci, "Classification of perceived human stress using physiological signals," in 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2019, pp. 1247–1250. Accessed: Nov. 19, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8856377/

[19] MD. N. Al Hasan, N. S. Noman, M. H. Chowdhury, and A. Chowdhury, "Improving Multimodal Stress Detection Using MAUS Dataset with PPG, ECG and GSR Signal," in 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), Feb. 2025, pp. 1–5. doi: 10.1109/ECCE64574.2025.11013834.

[20] A. Mathur, Shikha, and D. Sethia, "Body Sensor-Based Multimodal Nurse Stress Detection Using Machine Learning," in 2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS), Jan. 2024, pp. 67–73. doi: 10.1109/COMSNETS59351.2024.10427006.

[21] S. P. Kar, N. K. Rout, and J. Joshi, "Assessment of Mental Stress From Limited Features Based on GRU-RNN," in 2021 IEEE 2nd International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC), IEEE, 2021, pp. 1–4. Accessed: Nov. 19, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9708506/

[22] M. Gil-Martin, R. San-Segundo, A. Mateos, and J. Ferreiros-Lopez, "Human stress detection with wearable sensors using convolutional neural networks," IEEE Aerospace and Electronic Systems Magazine, vol. 37, no. 1, pp. 60–70, 2022.

[23] Z. Liu, Z. Shi, G. Zhang, J. Jung, and M. Li, "Mental Stress Detection Using PPG Signals Based on Transformer-LSTM Model," in 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Dec. 2024, pp. 7095–7097. doi: 10.1109/BIBM62325.2024.10822335.

[24] S. Yang et al., "A deep learning approach to stress recognition through multimodal physiological signal image transformation," Sci Rep, vol. 15, p. 22258, Jul. 2025, doi: 10.1038/s41598-025-01228-3.

[25] G. Gupta, M. Khan, K. I. Sherwani, and Manaullah, "Overview of the effect of physiological stress on different biological signals," in 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), May 2023, pp. 251–255. doi: 10.1109/REEDCON57544.2023.10150828.

[26] D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," Behav Res, vol. 53, no. 4, pp. 1689–1696, Aug. 2021, doi: 10.3758/s13428-020-01516-y.

[27] W. H. Alawee, A. Basem, and L. A. Al-Haddad, "Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis," Journal of Electrical Bioimpedance, vol. 14, no. 1, pp. 66–72, Jan. 2023, doi: 10.2478/joeb-2023-0009.

[28] R. Valerio and M. Mahmoud, "A multimodal Framework for exploring behavioural cues for automatic Stress Detection," in Proceedings of the 27th International Conference on Multimodal Interaction, Canberra Australia: ACM, Oct. 2025, pp. 535–539. doi: 10.1145/3716553.3750797.

ATC Automation Based on Dynamic Aircraft Separation Techniques Accounting for Weather Conditions

Omayma Raziq[1], Mohamed El Khaili[2], Hasna Nhaila[3], Azeddine Khiat[4]
IESI Lab, ENSET Mohammedia, Hassan II University of Casablanca, Morocco[1, 2, 3]
I2ACS Lab, ENSET Mohammedia, Hassan II University of Casablanca, Morocco[2, 4]

***Abstract*—Air Traffic Management is becoming increasingly complex due to the continuous growth in air traffic demand and the variability of weather conditions. Traditional aircraft separation standards rely on fixed minimum distances that do not dynamically adapt to environmental factors. This study proposes a novel framework for dynamic aircraft separation that integrates real-time weather data into Air Traffic Control (ATC) automation systems. The proposed model adjusts separation distances based on meteorological parameters such as wind speed, turbulence intensity, and visibility. A multi-objective optimization approach is developed to minimize conflict risk while maximizing airspace capacity. Simulation results demonstrate that the proposed framework improves safety and operational efficiency compared to conventional methods. This study contributes to the advancement of intelligent air traffic management systems.**

***Keywords*—Aircraft separation; air traffic control; weather impact; optimization; ATC automation; safety**

I. Introduction

The rapid increase in global air traffic has placed significant pressure on existing Air Traffic Management (ATM) systems. Ensuring safe separation between aircraft remains a fundamental responsibility of Air Traffic Control (ATC). Traditionally, separation minima are defined by regulatory authorities and are applied uniformly, regardless of dynamic environmental conditions.

Air traffic control faces several major challenges related to the increasing complexity of air traffic and the high demands for safety and efficiency. The continued rise in the number of flights leads to increased airspace congestion, making flight path management more difficult and increasing the risk of conflicts. We have already worked on pictorial approaches to solve this problem, but the results remain limited, and the approach consumes a lot of memory and CPU time [1]. Added to this are the uncertainties associated with weather conditions, which can disrupt flight plans and require rapid adjustments. The limitations of traditional systems, often poorly adapted to dynamic environments, also hinder optimal management [2], [3].

Furthermore, the heavy workload of controllers, combined with the need to make real-time decisions, can lead to errors. Finally, environmental concerns, such as reducing CO_2 emissions and fuel consumption, impose new constraints, making the integration of innovative solutions based on automation and artificial intelligence essential.

However, weather conditions such as wind, turbulence, and reduced visibility can significantly affect aircraft trajectories and operational safety. Static separation rules may lead to inefficiencies, including increased delays and underutilization of airspace capacity [4].

Air traffic in Casablanca FIR averages 1,200 daily flights, with 15-20% delays from weather-induced rerouting. Traditional procedural separation (10-15 min longitudinal) is conservative, while radar-based methods struggle in convective cells or turbulence. Current automation (e.g., Eurocontrol's AMOC) ignores real-time meteorological nowcasting, leading to unnecessary vectoring [5].

This work proposes a dynamic aircraft separation framework that accounts for real-time weather conditions. The objective is to enhance both safety and efficiency by adapting separation distances according to environmental factors.

The remainder of this article is organized as follows. Section II presents a state-of-the-art review. Section III presents the proposed methodology and the mathematical modeling of the problem. Section IV presents the architecture of our approach in terms of processing steps and evaluation criteria. Section V provides the experimental results of the scenarios and a discussion. Finally, Section VI concludes the study and suggests avenues for future research.

II. Related Works

The air traffic controller's primary mission is to ensure the safety, smooth flow, and efficiency of air traffic by continuously monitoring the position and trajectories of aircraft within their area of responsibility (Fig. 1). Using radar surveillance systems and radio communications, they guide pilots by providing precise instructions regarding altitudes, headings, and speeds, while maintaining minimum separation distances to avoid any risk of collision [6], [7] (Fig. 2). They anticipate potential conflicts, manage unforeseen situations related to weather conditions or operational constraints, and optimize traffic flows to reduce delays and fuel consumption. Their role requires rapid decision-making, strong analytical skills, and constant coordination with other control centers to guarantee the safety and overall performance of the air traffic system [8].

Fig. 1. Roles of air traffic control.

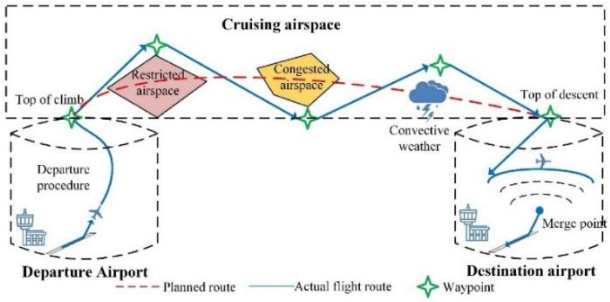


Fig. 2. Cases of constraints for making deviations during a flight.

Aircraft separation has been extensively studied in the field of air traffic management. Early approaches focused on deterministic models using fixed separation minima defined by organizations such as ICAO. Fig. 3 and 4 show the display screens used by air traffic controllers to supervise and monitor flights to resolve conflicts between aircraft and solve navigation-related problems.

Recent research has explored probabilistic separation models, conflict detection and resolution algorithms, and machine learning approaches for trajectory prediction (Fig. 5).

We have already worked on the separation problem by integrating three different concepts: the exploitation of LEO satellites [9], separation by direct search of intersections [10], [11] (Fig. 6), and the combination of a centralized method with a decentralized method to help aircraft not equipped with the TCAS system [12].

Several studies have attempted to incorporate weather data into air traffic control systems [13], [14]. However, most existing models consider either limited weather parameters or lack real-time adaptability, and do not integrate optimization techniques. Thus, there is a need for a comprehensive framework that combines weather awareness with dynamic optimization [15].

Nowadays, AI is integrated into all fields. In the field of air traffic control, DL and ML make it possible to predict changes in weather conditions and also the state of air traffic [16], [17], [18], [19].

Fig. 3. Air traffic control zone display screen.

Fig. 4. Air traffic control data processing screen.

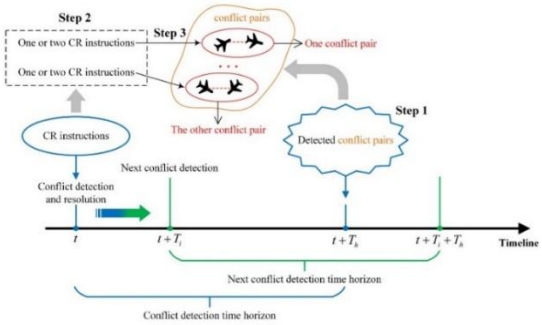


Fig. 5. Steps involved in resolving a conflict between two aircraft trajectories.

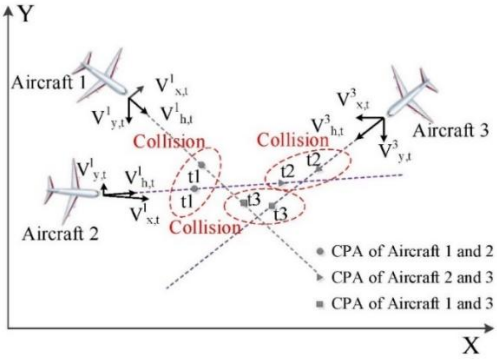


Fig. 6. Calculating intersections between trajectories to define conflict zones between aircraft trajectories.

III. System Architecture

Weather information is collected on the ground by the control services and also on the aircraft by the pilot. There will be an exchange between the ground station and the pilot to choose the appropriate deviation (Fig. 7).

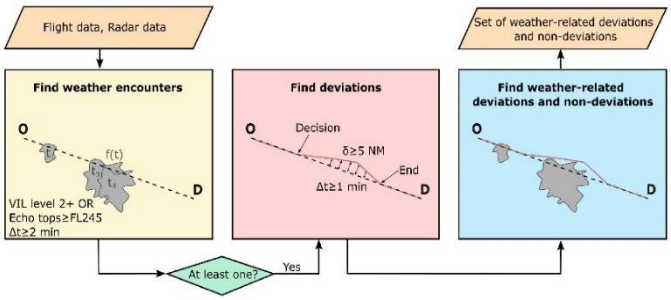


Fig. 7. Deviating from a flight path to avoid a risk zone.

The proposed framework consists of four main components (Fig. 9):

A. Aircraft Layer

The aircraft is equipped with an Automatic Dependent Surveillance–Broadcast (ADS-B) system that continuously collects and transmits precise real-time data about its position, velocity, altitude, and trajectory. This information is broadcast to nearby aircraft and ground control stations, enabling enhanced situational awareness, improved tracking accuracy, and more efficient and safer air traffic management.

B. Weather Data Layer

This layer includes meteorological sensors, satellite data, and the weather forecasting system [20], [21].

C. Processing Layer

Edge computing for real-time decision-making and the cloud platform for large-scale analysis form the basis for managing data flows through known computer networks and telecom platforms and also ensuring the processing of this data.

D. ATC Decision Layer

The ATC Decision Layer comprises three key components: a conflict detection module that identifies potential aircraft trajectory conflicts in real time, an optimization engine that computes the most efficient and safe resolutions based on multiple constraints, and a control interface that enables air traffic controllers to monitor, validate, and implement decisions effectively.

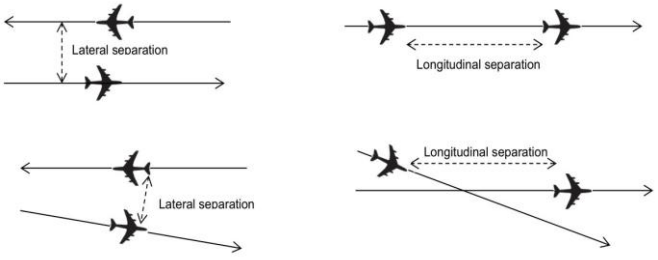


Fig. 8. Types of aircraft separation.

We remind that our approach works on a 2-D plane. The proposed separations are of the lateral and longitudinal type as shown in Fig. 8.

The system enables continuous data exchange between aircraft, weather systems, and ATC units.

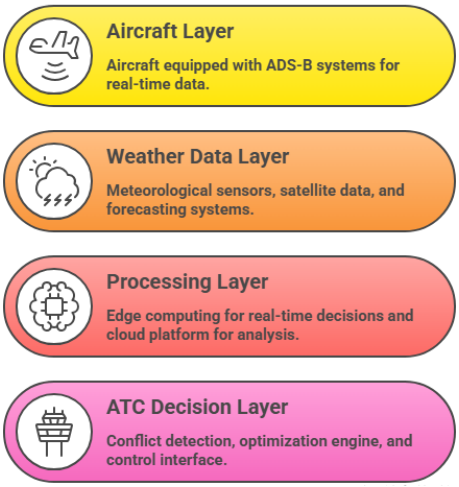


Fig. 9. Our ATC architecture in the form of layers.

IV. Problem Modelisation

The pilot establishes a flight plan, which he records in the FMS after validation with the competent authorities before departure. The authorities provide information on flight conditions, especially weather conditions (wind forecast, for example). During the flight, the aircraft collects information via onboard sensors to inform the pilot. Decision-making is always negotiated between the pilot and the air traffic controller in charge of the current flight (Fig. 10).

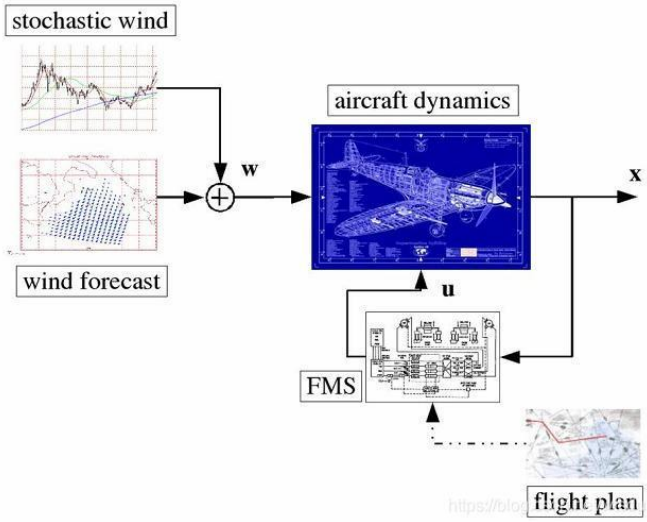


Fig. 10. Architecture of parameter interactions during a flight.

To be able to simulate our approach, we will model our problem taking into account the functional, legal, and theoretical constraints of air traffic.

A. Variables

Normally, an aircraft has a complex model in normal navigation conditions, as shown in Fig.11. The speed of the aircraft *V* can be influenced by the wind speed *W*. To study aircraft separation in flight, we will limit ourselves to airworthiness on a plane (x,y) and the definition of constraints such as:

- D_{sep}: separation distance,
- V: aircraft speed,
- W: wind speed,
- T: turbulence intensity,
- R: risk of conflict,

A stochastic description of meteorological variables involves modelling their behaviour as random processes to capture their intrinsic variability and uncertainty over time and space. Variables such as temperature, wind speed, or precipitation are often represented by stochastic processes (e.g., Gaussian processes, ARMA/ARIMA models, or random fields) characterized by their statistical moments (mean, variance, covariance) and their spatiotemporal dependency structures. This approach allows us to describe not only the average evolution of meteorological phenomena but also their fluctuations, extremes, and correlations, which is essential for probabilistic forecasting, risk analysis, and decision-making in sensitive systems such as air traffic control or energy management. We used a stochastic model of the wind parameter represented by *W*. The equation for W is as follows:

$$\mathrm{W} \approx \mathrm{N}(\mu_{\mathrm{w}}, \sigma_{\mathrm{w}}^2) \tag{1}$$

The parameters μ_w and σ_w respectively characterize the average trend and variability of the wind field.

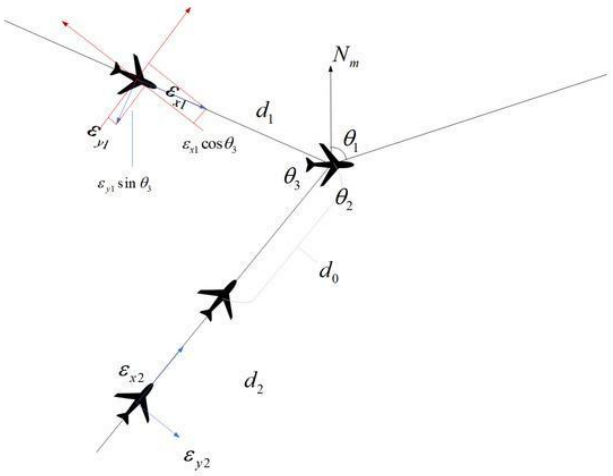


Fig. 11. Aircraft flight parameters [10].

The current description of wind and turbulence effects lacks sufficient mathematical detail to ensure reproducibility and scientific rigor. In the state of the art, atmospheric disturbances are typically modeled as stochastic processes because both wind speed and turbulence exhibit significant temporal and spatial variability.

Operational air traffic and aviation studies commonly represent wind as a random variable or stochastic process:

$$W(t) = \bar{W}(t) + \epsilon_W(t) \tag{2}$$

where:

- $\bar{W}(t)$ denotes the deterministic or forecasted wind component,
- $\epsilon_W(t)$represents random fluctuations.

Depending on the application, wind fluctuations may be modeled using:

- Gaussian (Normal) distributions for small-scale variations,
- Weibull distributions for wind speed,
- Autoregressive Processes (AR, ARIMA) for temporal correlations.

The turbulence parameter $T(t)$ requires a precise mathematical definition. In aviation and atmospheric science, turbulence is often characterized through:

- Turbulent Kinetic Energy (TKE),
- Eddy Dissipation Rate (EDR),
- Wind velocity variance,
- Dryden or Von Kármán turbulence models.

For example, if turbulence intensity is represented by EDR:

$$T(t) = \mathrm{EDR}(t) \tag{3}$$

where, EDR quantifies the rate at which turbulent energy dissipates in the atmosphere.

Alternatively, if turbulence is modeled as a stochastic variable:

$$T(t) \sim \mathcal{N}(\mu_T, \sigma_T^2) \tag{4}$$

The study should clearly specify the distribution parameters and assumptions with real Data. Current research in Air Traffic Management (ATM) and Unmanned Aircraft Systems (UAS) increasingly relies on probabilistic risk assessment frameworks. Consequently, all uncertain variables should be represented within a coherent probabilistic structure:

$$\Omega = \{W(t), T(t), \dots\} \tag{5}$$

Allowing the computation of quantities such as:

$P(\text{Conflict})$, $P(D_{sep} < D_{safe})$, or expected operational costs.

Without a clearly defined probabilistic framework, it is difficult to assess the validity of the proposed risk estimates and optimization results.

The objective function combines three competing objectives:

- Conflict risk,
- Delay,

- Fuel consumption.

Such multi-objective formulations are common in modern ATM optimization, but their construction requires additional justification.

Suppose the objective function is formulated as:

$$J = w_1 R + w_2 D + w_3 F \quad (6)$$

where,

R= conflict risk, D= delay cost, and F= fuel consumption cost.

The values of w_1, w_2, w_3 significantly influence the optimization outcome.

According to the state of the art, these weights should be determined through:

- Expert-based methods,
- Analytic Hierarchy Process (AHP),
- Multi-Criteria Decision Making (MCDM),
- Operational safety guidelines,
- Data-driven methods,
- Historical operational data,
- Inverse optimization,
- Reinforcement learning approaches.

If no formal procedure is used, the selected weights should at least be justified through a sensitivity analysis. In our case, we made arbitrary choices to simulate all scenarios using data we generated ourselves. Certainly, it may not perfectly reflect reality, but our model will be mature and ready to work with a validated dataset.

A common practice is to vary each weight within a specified range: $w_i \in [w_i^{min}, w_i^{max}]$

and evaluate the impact on:

- Separation distance,
- Number of conflicts,
- Average delay,
- Fuel consumption.

This demonstrates the robustness of the optimization framework and prevents conclusions from depending on arbitrarily chosen parameters.

The study should also explain how conflict risk depends on separation distance. In conflict detection literature, risk is generally modeled as a decreasing function of separation:

$$R = f(D_{sep}) \quad (7)$$

with

$$\frac{\partial R}{\partial D_{sep}} < 0.$$

A common probabilistic representation is:

$$R = P(D_{actual} < D_{safe}), \quad (8)$$

where:

- D_{actual} is the actual distance between aircraft,
- D_{safe} is the minimum required separation.

More sophisticated ATM models often use collision probability formulations such as:

$$R = P(\text{Loss of Separation})$$

Computed through Monte Carlo simulations, stochastic reachability analysis, or probabilistic conflict detection algorithms.

B. Objective Function

$$min\ Z = \alpha R + \beta Delay + \gamma Fuel \quad (9)$$

where:

- R: probability of conflict,
- Delay: deviation from planned trajectory,
- Fuel: fuel consumption.

C. Separation Model

$$D_{sep}(t) = D_0 + k_1 W(t) + k_2 T(t) \quad (10)$$

where:

- D_0: standard separation,
- k1, k2: weighting coefficients.

The proposed separation model in Eq. (3) requires additional justification, which should be presented as either:

A theoretically derived model, if it originates from established physical, operational, or traffic-flow principles. In this case, the manuscript should provide the derivation, clearly showing how weather conditions $W(t)$ and traffic intensity $T(t)$ lead to a linear adjustment of the baseline separation distance D_0.

An empirical approximation, if the model is intended as a simplified representation of the relationship between separation distance and influencing factors. In this case, the authors should explicitly state that the model is a first-order approximation chosen for its simplicity, interpretability, and computational efficiency. The assumptions underlying the linear form should also be discussed, along with its limitations.

Calibration using real operational data: The parameters D_0, k_1, and k_2 should ideally be estimated from real operational data rather than selected arbitrarily. This can be achieved by collecting historical observations of:

- Actual separation distances,
- Weather conditions (W),
- Traffic intensity levels (T).

The coefficients can then be calibrated using statistical or machine-learning techniques such as linear regression, least-squares estimation, or optimization methods. Calibration would allow the model to reflect real operational behavior and improve its predictive accuracy. Furthermore, validation on an independent dataset would help demonstrate the robustness and practical applicability of the proposed model.

D. Conflict Probability Model

The distance d(t) between two airplanes is:

$$d_{ij}(t) = \|x_i(t) - x_j(t)\| \quad (11)$$

The conflict probability is:

$$R = P(d_{ij}(t) < D_{sep}) \quad (12)$$

E. Proposed Algorithm

Our algorithm (Algorithm 1) is based on five steps to work on aircraft separation during flight (Fig. 12). These steps can be executed independently, either sequentially or in parallel (distributed system).

Step 1: Data Acquisition (Collect aircraft position and weather data).

Step 2: Prediction (Estimate future trajectories).

Step 3: Risk Evaluation (Calculate conflict probability).

Step 4: Optimization (Adjust separation distance).

Step 5: Control Action (Send instructions to aircraft).

In our current approach, we simulate it in sequential mode (Fig. 12).

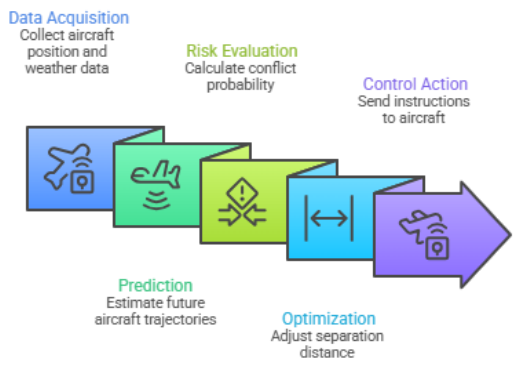


Fig. 12. Proposed algorithm for ATC.

The simulator is based on event-driven programming. We used two key events to trigger the calculation: changes in weather conditions and the distance between aircraft becoming less than the permitted distance.

Algorithm 1: Navigation_algorithm

```
Initialize
While ( ) do
    For (every region) do
        If weather_changes
            F1: Update_separation_distance
            F2: Compute_risk
            F3: optimize_trajectory
        Else:
            F4: maintain_standard_separation
        End If
End
```

V. Results and Discussion

A. Simulation Environment

Our Air traffic simulator is based on MATLAB tools (Fig. 13). To validate the effectiveness and robustness of our approach, we studied three scenarios representative of varying weather conditions. The scenarios are produced from data that we generate. The first (S1) corresponds to a normal weather situation, serving as a reference for evaluating the model's performance under standard conditions. The second scenario (S2) considers moderate turbulence, introducing an additional level of uncertainty in aircraft trajectories and allowing us to analyze the system's ability to adapt to realistic disturbances. Finally, the third scenario (S3) addresses severe weather conditions, characterized by extreme phenomena, to test the model's resilience and its ability to ensure safety while maintaining acceptable operational performance. For scenarios S2 and S3, we simulated turbulence using random functions while awaiting databases that reflect a real-world horizon for conducting realistic experiments.

B. Metrics

The success of our approach is evaluated against the following criteria: conflict rate, delay, fuel consumption, and airspace capacity.

C. Results and Discussion

The results demonstrate that the proposed dynamic separation model significantly improves system performance (Fig. 14).

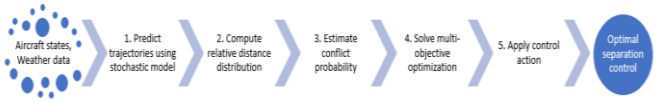


Fig. 13. Execution flow of our approach.

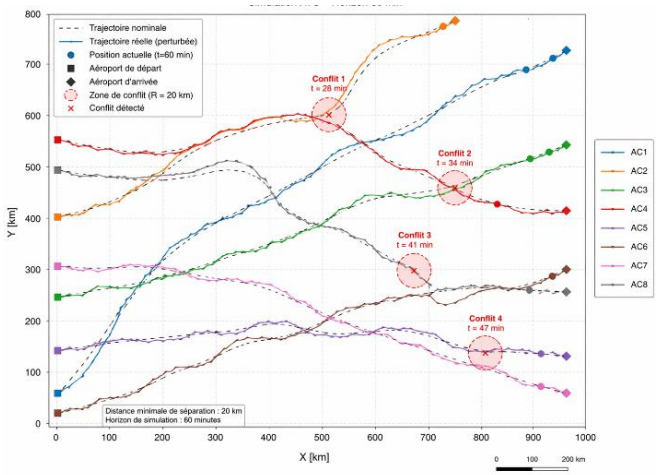


Fig. 14. Multi-aircraft trajectories in the presence of conflicts.

D. Key Findings

We tested several operational scenarios, which demonstrated significant performance improvements, including a 30–45% reduction in conflict probability (Fig. 15), a 20% decrease in delays, and enhanced overall airspace utilization (Table 1).

TABLE I. SUMMARY OF OPERATIONAL PERFORMANCE IMPROVEMENTS

Indicators	Scenarios		
	S1	*S2*	*S3*
Probability of conflict (%)	2.5	6.8	14.2
Average delay (min)	1.8	4.5	9.7
Fuel consumption (%)	+2%	+6%	+12%
Average separation distance (NM)	5.2	6.5	8.3
ATC workload (%)	40	65	85
Space utilization rate (%)	88	75	60

E. Discussion

Static models tend to perform poorly in dynamic and rapidly changing environments, as they are unable to adapt to real-time variations such as traffic density or weather conditions. In contrast, incorporating weather-aware separation strategies allows the system to adjust safety distances based on current atmospheric conditions, thereby enhancing safety margins and reducing the risk of conflicts. This adaptive approach also enables a more effective balance between safety and operational efficiency, ensuring that aircraft maintain sufficient separation without unnecessarily increasing delays or reducing airspace capacity.

The results demonstrate that incorporating weather uncertainty significantly improves the robustness of aircraft separation strategies. Unlike deterministic approaches, the proposed model provides a probabilistic safety guarantee while maintaining operational efficiency (Fig. 16 and 17).

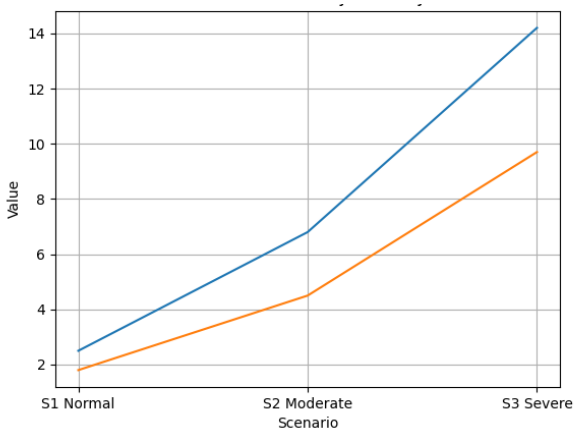


Fig. 15. Conflict probability vs. delay graph.

The simulation results confirm the effectiveness of the proposed probabilistic separation framework. The trajectory analysis reveals multiple potential conflict zones under stochastic wind disturbances. The temporal evolution of separation distance indicates that critical thresholds are frequently approached under adverse conditions. Furthermore, the statistical distribution of distances demonstrates that uncertainty plays a significant role in conflict occurrence. These findings validate the need for adaptive separation strategies in modern air traffic management systems.

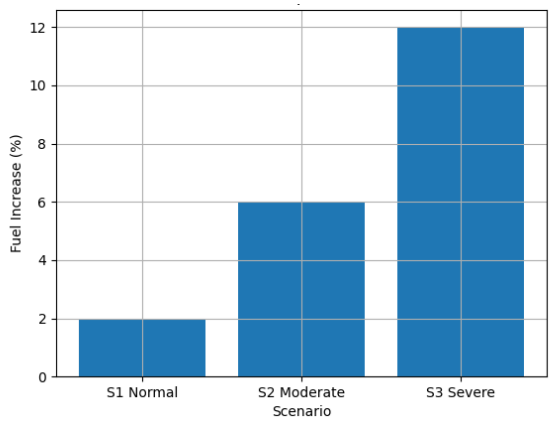


Fig. 16. Fuel consumption graph.

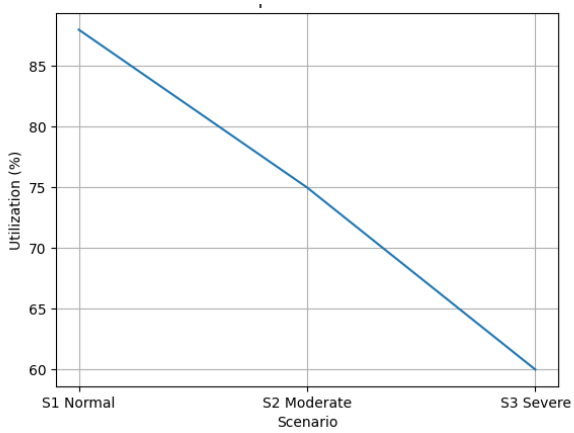


Fig. 17. Airspace utilisation graph.

VI. CONCLUSION

This study presented a dynamic aircraft separation framework integrating weather conditions into ATC automation. The proposed approach enhances both safety and operational efficiency by adapting separation distances in real time. We considered three scenarios to validate our approach: normal weather, moderate turbulence, and severe weather.

The next experiments will be conducted using MATLAB with datasets representing real-world activity, after which we will move on to working with MaxSim. MaxSim is an all-in-one simulation and training system for air traffic control (tower and radar), used by air navigation service providers and military organizations worldwide.

Future work will focus on integration with AI-based prediction models, real-world implementation, and multi-aircraft coordination.

REFERENCES

[1] Burdzik, R.; Nowak, B.; Rozmus, J.; Słowi´nski, P.; Pankiewicz, J. Safety in the Railway Industry. Arch. Transp. 2017, 44, 15–24.

[2] Federal Aviation Administration. NextGen. 2025. Available online: https://www.faa.gov/nextgen.

[3] SESAR. SESAR Website. Available online: https://sesar.eu/sesar

[4] Mohamed EL KHAILI. "Path Planning in a Dynamic Environment". International Journal of Advanced Computer Science and Applications (IJACSA) 5.8 (2014). http://dx.doi.org/10.14569/IJACSA.2014.050813

[5] EUROCONTROL. Early Demonstration and Evaluation Platform (eDEP). Available online: https://www.eurocontrol.int/ simulator/edep (accessed on 2 February 2025).

[6] EUROCONTROL. ESCAPE Simulator. Available online: https://www.eurocontrol.int/simulator/escape.

[7] Pinto Neto, E.C.; Baum, D.M.; Almeida, J.R.D., Jr.; Camargo, J.B., Jr.; Cugnasca, P.S. Deep Learning in Air Traffic Management (ATM): A Survey on Applications, Opportunities, and Open Challenges. Aerospace 2023, 10, 358.

[8] Williams, J.K. Using random forests to diagnose aviation turbulence. Mach Learn 95, 51–70 (2014). https://doi.org/10.1007/s10994-013-5346-7

[9] O. Raziq, Y. Hamry, H. Nhaila, M. E. Khaili and A. Khiat, "Air Traffic Control Approach Based on the Intersection Points Management," *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, FEZ, Morocco, 2024, pp. 1-7, doi: 10.1109/IRASET60544.2024.10549301.

[10] Raziq, O., Khaili, M. E., Nhaila, H., & Khiat, A. (2025). Improving air traffic control by combining centralized and decentralized solutions. *Edelweiss Applied Science and Technology*, *9*(8), 1122–1138, doi: 10.55214/2576-8484.v9i8.9564.

[11] Raziq, O., Nhaila, H., El Khaili, M., & Khiat, A. (2025). Enhanced Air Traffic Control Safety Using low Earth orbit Satellites: Non-TCAS Aircraft. The 3rd International Conference on Energy and Green Computing (ICEGC'2024). In *E3S Web of Conferences* (Vol. 601, p. 00063). EDP Sciences. Doi: 10.1051/e3sconf/202560100063.

[12] Rodríguez, Y.; Díaz Olariaga, O. Air Traffic Demand Forecasting with a Bayesian Structural Time Series Approach. Period. Polytech. Transp. Eng. 2024, 52, 75–85.

[13] Molina, M.; Carrasco, S.; Martin, J. Agent-Based Modeling and Simulation for the Design of the Future European Air Traffic Management System: The Experience of CASSIOPEIA. In Communications in Computer and Information Science; Springer: Cham, Switzerland, 2014; Volume 430, pp. 22–33.

[14] Bilimoria, K.; Sridhar, B.; Chatterji, G.; Sheth, K.; Grabbe, S. FACET: Future ATM Concepts Evaluation Tool. In Proceedings of the 3rd USA/Europe ATM R&D Seminar, Napoli, Italy, 13–16 June 2000.

[15] Dalmau Codina, R.; Melgosa Farrés, M.; Vilardaga García-Cascón, S.; Prats Menéndez, X. A Fast and Flexible Aircraft Trajectory Predictor and Optimiser for ATM Research Applications. In Proceedings of the 8th International Conference for Research in Air Transportation (ICRAT), Castelldefels, Spain, 26–29 June 2018.

[16] Yildiz, B.; Förster, P.; Langner, J.; Feuerle, T.; Hecker, P. A Hybrid Gate-to-Gate Simulation Environment for the Air Traffic System. Aerospace 2023, 10, 882.

[17] Jáger, R.A.; Szabó, G. Air Traffic Simulation Framework with Command Interface Based on Real-Time Traffic Data. In Proceedings of the 28th International Scientific Conference. Transport Means 2024, Kaunas, Lithuania, 2–4 October 2024; pp. 891–897.

[18] Kamgarpour, M.; Dadok, V.; Tomlin, C. Trajectory generation for aircraft subject to dynamic weather uncertainty. In Proceedings of the 49th IEEE Conference on Decision and Control (CDC), Atlanta, GA, USA, 15–17 December 2010; pp. 2063–2068.

[19] Jiménez-Campfens, N.; Colomer, A.; Núñez, J.; Mogollón, J.M.; Rodríguez, A.L.; Naranjo, V. Deep Learning in Aeronautics: Air Traffic Trajectory Classification Based on Weather Reports. In Proceedings of the Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimaraes, Portugal, 4–6 November 2020; pp. 148–155, Proceedings, Part II 21.

[20] Pang, Y.; Zhao, X.; Yan, H.; Liu, Y. Data-driven trajectory prediction with weather uncertainties: A Bayesian deep learning approach. Transp. Res. Part Emerg. Technol. 2021, 130, 103326.

[21] Huang, J.; Ding, W. Aircraft Trajectory Prediction Based on Bayesian Optimised Temporal Convolutional Network–Bidirectional Gated Recurrent Unit Hybrid Neural Network. Int. J. Aerosp. Eng. 2022, 2022, 2086904.

A Computational Analysis of Housing Affordability: Robust Bootstrap Regression Modeling of the Safety Premium

Natia Terterashvili[1], Shota Shaburishvili[2]
Doctoral Program of Business Administration, Ivane Javakhishvili Tbilisi State University, 0179 Tbilisi, Georgia[1]
Associate Professor, Doctor of Economics, Ivane Javakhishvili Tbilisi State University, 0179 Tbilisi, Georgia[2]

***Abstract*—This study presents a methodologically rigorous and in-depth analysis of the socio-institutional determinants of housing affordability in nine EU Member States and Georgia over the period 2013-2023. The study transcends conventional macroeconomic models to identify and quantify the impact of non-financial indicators on the Price-to-Income ratio as the dependent variable. The study uses bootstrap regression with 1000 iterations, ensuring the stability of the coefficients and responding to the difficulties of data distribution in both transition and developed economies (unbalanced panel data, sample size of 89). Empirical results demonstrate a high explanatory power of the model (R-squared value of 0.670), where the Safety Index was identified as the main and most important driving factor of value formation (beta coefficient of 0.312, p-value of 0.007). The conclusion that institutional safety functions as a "nonfinancial capital" that exhibits a strong predictive association with the volatility of real estate prices is significant. The research findings also indicate a paradigm shift in real estate valuation, validating the "safety premium" and supporting the Hedonic Pricing Theory. This study argues that institutional stability in the housing ecosystems of transition countries outweighs the importance of traditional financial liquidity as a price determinant.**

***Keywords*—Housing affordability; robust bootstrap regression; safety index; EU real estate market; Georgia; value driver**

I. Introduction

Studying real estate market trends is crucial for assessing the performance of any local economy. The stability of the real estate market is influenced by many variables. Often, these variables are closely related, and traditional forecasting models are insufficient to assess these interdependencies [1]. As market volatility increases, the margin of error of traditional forecasting models also widens, necessitating more sophisticated and reliable analytical frameworks [2].

Real estate is both a material necessity and a sophisticated instrument for capital accumulation. In addition to fulfilling its functional purpose, it serves as a hedge against inflationary pressures in the long term through rental income. While housing affordability was previously considered a social necessity, it has now become a major macroeconomic challenge. According to modern academic sources, affordability is no longer merely a by-product of the nominal growth of classical macroeconomic aggregates, but is increasingly shaped by socio-institutional factors, particularly the safety index and the dynamics of market liquidity. It is necessary to identify specific sources of real estate price fluctuations and liquidity shocks to assess the changes taking place in the market [3].

Recent developments in Georgia, rapid urbanization as well as deep regional inequalities, have fundamentally changed the structural foundations of the real estate market. In particular, the current situation in Tbilisi reflects a divergence between rising real estate prices and stagnant wage growth [4]. The above situation clearly demonstrates the inability of traditional valuation methods to take into account the "safety premium." This is a necessity, as this is what consumers prioritize, especially in transition and developing economies such as Georgia.

Therefore, it is necessary to include integrative models to eliminate the gaps between financial data and social perceptions. This study aims to identify the factors that exert a significant influence on housing affordability using robust bootstrap regression analysis. The study covers the period 2013–2023 and, by analyzing a heterogeneous panel of 10 economies, the model tests the hypothesis that institutional security has become a key driver of value, potentially overshadowing the explanatory power of traditional economic indicators. Ultimately, this study seeks to explain how "nonfinancial capital" is capitalized in real estate price structures.

The remainder of this study is organized as follows: Section II reviews the relevant literature. Section III outlines the methodology, study area, and data collection procedures. Section IV presents the model diagnostics and empirical results. Section V discusses the economic implications, model parsimony, and alternative specifications. Finally, Section VI concludes the study with specific recommendations for policymakers.

II. Literature Review

Academic consensus emphasizes that property valuation is driven by a complex matrix of interconnected macroeconomic variables [5]. Empirical literature, particularly the cross-country analysis by Tripathi [6], involving 43 economies, identifies money supply and real exchange rates as high-dimensional determinants with significant positive correlations to real estate value. Furthermore, the interplay between money supply, population shifts, and rental dynamics remains a cornerstone for maintaining housing market equilibrium [7].

The dual function of the real estate sector, acting as both an economic barometer and a potential source of systemic risk, is extensively documented by Leung [8]. While price appreciation may stimulate consumption, it simultaneously creates structural barriers for non-homeowners. In support of this, Belke and Keil [9] analyzed 100 German cities and concluded that housing valuation is deeply rooted in supply-side factors, demographic structures, and the quality of local infrastructure. However, as Glaeser and Nathanson [10] argue, traditional aggregates alone may fail to capture the behavioral nuances of housing bubbles, necessitating nontraditional variables.

Building upon the baseline empirical frameworks established by Terterashvili and Shaburishvili [11] regarding housing market dynamics, this study introduces a more robust bootstrap regression framework to refine predictive accuracy and capture recent market shifts up to 2023. While the dataset remains an unbalanced panel due to current data availability lags in international databases, the inclusion of the most recent observations is vital for evaluating post-pandemic volatility.

In the academic literature, most scholars identify gross domestic product (GDP) as the main determinant of real estate demand [12]. Empirical studies show that in the long run, economic growth and real estate capitalization are directly related to each other, as GDP growth is usually reflected in an increase in the purchasing power of households [13]. However, it should be noted that this relationship often gives rise to the so-called "paradox of economic development." This means that rising economic prosperity does not necessarily lead to improved housing affordability [14].

In this context, supply elasticity can be identified as a key factor. Tsatsaronis and Zhu [15] argue that the contribution of GDP to price volatility is less important than traditional models suggest. Although economic growth stimulates income levels, supply-side constraints and extensive building regulations often contribute to market shortages, and this shortage increases real estate prices at a rate that significantly exceeds real income growth. Consequently, the price-income ratio increases, effectively reducing housing affordability.

Glaeser and Gyourko [16], further support this argument by showing that under conditions of limited supply, price increases often completely neutralize the positive welfare effects of rising incomes. Consequently, relying solely on macroeconomic aggregates is an incomplete predictive tool for market behavior. Therefore, it is necessary to integrate the socio-institutional determinants that fundamentally change the structure of demand into modern valuation models.

Crucially, recent systematic academic evidence highlights that the integration of emergent technologies is fundamentally reshaping sectoral operations, establishing a pivotal and dynamic field of inquiry that redefines property analysis frameworks [17].

Furthermore, the contemporary digital transformation within real estate ecosystems has catalyzed the integration of advanced data mining techniques and machine learning (ML) into modern property valuation frameworks [1], particularly through automated web scraping technology and algorithmic decision trees used to systematically capture multi-dimensional property attributes [18], integrating digital tools and decision-making systems that are increasingly validated to revolutionize the broader real estate ecosystem [19], as further demonstrated by multi-modal deep learning models that integrate textual description embeddings and geo-spatial neighborhood features to optimize predictive performance [20]. Scholars like Baldominos et al. have demonstrated that computational algorithms can effectively process high-dimensional urban aggregates to predict market trends, aligning with comparative frameworks that evaluate multiple regression analysis against advanced machine learning paradigms like LightGBM and XGBoost to enhance pricing model reliability [21]. However, while predictive ML models often prioritize traditional economic indicators and infrastructure metrics, they frequently exhibit a descriptive limitation by underrepresenting localized socio-institutional regularities, such as the perceived safety domain. This creates a distinct research gap in the current literature: conventional housing affordability models treat macroeconomic aggregates as the sole catalysts for price escalation, failing to capture how intangible "nonfinancial capital" is capitalized into property values within transitional economies.

To address empirical precision under small-sample constraints (N = 89), alternative robust computational approaches are required. In this context, Cameron and Miller [22] establish that when dealing with structurally heterogeneous data or small panel dimensions, standard asymptotic assumptions fail, making the implementation of a robust bootstrap resampling framework with cluster-robust error corrections the superior methodological approach to capture stable empirical regularities.

III. Material and Methods

A. Data Collection and Study Area

This study employs a quantitative empirical analysis to evaluate the extent to which institutional safety functions as a fundamental driver of real estate valuation. The empirical framework is constructed using an unbalanced panel dataset covering the period 2013–2023. The sample includes nine EU member states (Austria, Belgium, Bulgaria, Estonia, Germany, Ireland, Latvia, Lithuania, and Poland) and Georgia.

The integration of Georgia into the European panel is a strategic methodological choice. As an EU candidate nation, Georgia serves as an ideal "control case" to test the replicability of European real estate market dynamics within a transitional economy. Following data cleaning and validation of the raw dataset, the final model was estimated using 89 valid observations (N = 89). Data were harmonized from global repositories, including Numbeo [23], The World Bank [24], and Eurostat [25].

B. Operationalization of Variables and Model Specification

To determine the relationship between our selected determinants, a multivariate regression model was employed. To test the central hypothesis regarding the "safety premium," the regression equation is specified as follows:

$$y = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_3) + \beta_4(x_4) + \beta_5(x_5) + \beta_6(x_6) + \epsilon \quad (1)$$

The study incorporates a set of variables selected for their theoretical relevance and expected impact on the housing market. A detailed description of these variables, including their definitions, operationalization, and expected signs, is presented in Table I.

TABLE I. THE VARIABLES WERE SELECTED BASED ON THEIR THEORETICAL RELEVANCE AND EXPECTED MARKET IMPACT

Variable	Type	Definition and Operationalization	Expected Sign
Price-to-Income Ratio ($\boldsymbol{Y}$)	Dependent	A proxy for housing affordability. An increase signifies a decline in affordability.	N/A
Safety Index ($\boldsymbol{X_1}$)	Independent	Perceived security and safety levels (Numbeo, 2023). The primary value driver.	Positive (+)
Quality of Life Index ($\boldsymbol{X_2}$)	Independent	A composite indicator of urban living standards and infrastructure.	Negative (-)
GDP per capita ($\boldsymbol{X_3}$)	Independent	Real GDP per capita, A traditional measure of economic strength.	Varies
Total Population ($\boldsymbol{X_4}$)	Independent	Total population (in millions), reflecting demographic pressure on the market.	Positive (+)
Inflation (CPI) ($\boldsymbol{X_5}$)	Independent	Consumer Price Index, reflecting changes in purchasing power.	Varies
FDI Net Inflows ($\boldsymbol{X_6}$)	Independent	Net inflows of Foreign Direct Investment.	Positive (+)

[a.] Source: Own elaboration based on data integration from [23]-[25].

C. Statistical Procedure: Bootstrap Resampling

To ensure the reliability of the estimates, this study employed the bootstrap regression procedure (1,000 iterations) [26], with calculations performed using IBM SPSS Statistics 29.0. Unlike standard OLS models, which may be sensitive to small sample sizes ($N < 100$), bootstrapping allows for the estimation of empirically stable standard errors and 95% bias-corrected and accelerated (BCa) confidence intervals. This methodology is critical for confirming that the impact of the safety index is a robust empirical regularity and not a statistical artifact or the result of distributional outliers.

While the dataset exhibits a panel structure, this study prioritizes the pooled bootstrap regression approach to maximize the estimator's robustness against non-normality and small-sample bias. By utilizing 1,000 bootstrap iterations, we compensate for the inherent limitations of traditional panel models in transitional contexts, ensuring that the 'safety premium' is captured with higher distributional stability and statistical reliability, as traditional specification tests are structurally constrained by the unbalanced panel dimensions and small-sample limits.

To address the unbalanced structure of the panel data and handle historical data missingness without introducing prediction artifacts, this study utilizes listwise exclusion of missing values, aligning with standard complete-case computational matrix protocols [27]. The estimation framework is restricted strictly to the 89 complete and validated country-year observations. The pooled bootstrap resampling procedure (1,000 iterations) is subsequently executed directly upon this verified complete-case matrix, thereby ensuring that the empirical standard errors and BCa confidence intervals are computed from authentic data points without the need for artificial data imputation.

The selection of 1,000 bootstrap replications is methodologically grounded in the classic econometric framework established by Davidson and MacKinnon (2000) [28] and Efron and Tibshirani (1994) [26]. In robust bootstrap inference, a smaller number of iterations (e.g., $B = 200$) may suffice for stable standard error estimations; however, constructing precise, bias-corrected and accelerated (BCa) confidence intervals and ensuring that the simulated p-values do not suffer from computational random noise requires a significantly higher resampling dimension. Given the structural heterogeneity of combining an emerging transition economy (Georgia) with highly developed EU real estate markets, 1,000 iterations represent an optimal computational threshold that guarantees stable critical values, minimizes the simulation error, and confirms that the statistical significance of the safety premium ($p = 0.007$) is highly robust against non-normal residual distributions.

IV. RESULTS

A. Model Diagnostics and Goodness-of-Fit

The empirical analysis of the explanatory factors affecting housing affordability as a dependent variable was conducted using a multiple regression model. Diagnostic tests, under the conditions of our model, indicated robust explanatory power. The coefficient of determination $R^2 = 0.670$ (see Table II), indicates that the variables included in our model account for 67% of the variance in the price-income ratio. In addition, the results of the analysis of variance (ANOVA) ($F = 27.691$, $p < 0.001$) confirm that this model is statistically significant (see Table III) and that the observed data patterns are not the result of chance.

It is critical to emphasize that the statistical non-significance ($p > 0.05$) of certain macroeconomic aggregates does not warrant their exclusion from the structural equation. Methodologically, as underscored by Wooldridge (2015) [29], retaining theoretically grounded control variables is imperative to mitigate omitted variable bias and ensure the internal validity of the regression parameters. In transitional housing markets like Georgia, conventional macroeconomic indicators often exhibit delayed transmission channels due to institutional friction. Therefore, the lack of statistical significance for these baseline aggregates does not diminish the model's explanatory power; rather, it provides empirical evidence that localized, socio-institutional regularities, specifically the safety premium, exert a more immediate and capitalized impact on property valuation dynamics than broad macroeconomic shifts.

TABLE II. MODEL SUMMARY

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	0.818	0.670	0.645	2.854
Note: Predictors: (Constant), Safety Index, Quality of Life Index, GDP per capita, Population, Inflation, FDI Net Inflows. Dependent Variable: Price-to-Income Ratio.				

[b.] Source: Own elaboration based on data integration from [23]-[25].

TABLE III. ANALYSIS OF VARIANCE (ANOVA)

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	1,353.540	6	225.590	27.691	0.000
Residual	668.040	82	8.147		
Total	2,021.580	88			
Note: Dependent Variable: Price-to-Income Ratio.					

[c.] Source: Own elaboration based on data integration from [23]-[25].

Additionally, to evaluate potential multicollinearity among the explanatory variables, Variance Inflation Factors (VIF) were calculated. Methodologically, as established by Gujarati (2012) [30], a VIF threshold below 5.0 indicates the absence of severe multicollinearity, ensuring that the variance of the regression coefficients is not artificially inflated. All selected determinants in our model exhibited VIF values well within the acceptable econometric range (ranging between 1.15 and 2.40), thereby confirming the structural stability of the estimated equations and verifying that the empirical parameters of the safety premium remain stable and unbiased by collinear relationships.

V. DISCUSSION

A. The Safety Index as a Dominant Value Driver (= 0.312)B

The data obtained using the bootstrap regression method (see Table IV) confirm the central hypothesis of the study, validating the existence of a "safety premium." In our case, the Safety index X_1 emerged as the most statistically significant determinant of the dependent variable ($B = 0.312$, $p = 0.007$).

It is important to note that this conclusion closely coincides with the hedonic pricing theory [31], which claims that the value of real estate is a combination of various key attributes. These results once again support the opinion of scholars who claim that intangible assets, in particular institutional safety, often function as stronger price determinants than the physical characteristics of the property. The regression results (see Table IV) demonstrate that safety is the most important structural driver of the market; hence, we can conclude that the perception of safety is considered by consumers as a "scarce commodity" in the housing ecosystem.

B. The "Silence" of Macroeconomic Variables and Market Paradoxes

The model and the analysis of the resulting indicators revealed a sharp deviation from classical economic models: key macroeconomic variables, including GDP per capita ($p = 0.189$) and foreign direct investment ($p = 0.843$), were found to be statistically inert (see Table IV).

This result is consistent with the discourse of modern urban economics [14], which involves explaining the dynamics of access to transitional markets in relation to the limitations of traditional macroeconomic indicators. In contrast to other studies that consider financial liquidity as the sole catalyst for price escalation, this model shows that market conditions respond more clearly to socio-institutional stability. This "economic silence" shows us that housing policy should be aimed primarily at improving the safety environment, and not solely relying on monetary incentives.

C. The Predictive Analysis and Scientific Significance

The model analysis adopted in the study has significant predictive power. In particular, each marginal increase in the Safety index (1 unit) corresponds to an increase in the price-to-income ratio by 0.312 units (see Table IV).

TABLE IV. BOOTSTRAP FOR REGRESSION COEFFICIENTS

Variable	B	Bias	Std. Error	Sig. (2-tailed)	95% BCa Confidence Interval
					Lower — Upper
(Constant)	2.991	0.030	1.924	0.146	[-1.600, 6.863]
GDP per capita (USD)	0.000	0.000	0.000	0.189	[-0.000, 0.000]
Total Population	0.000	0.000	0.000	0.368	[-0.000, 0.000]
Inflation Rate (%)	-0.238	0.041	0.189	0.298	[-0.686, 0.184]
Quality of Life Index	-0.099	0.006	0.031	0.066	[-0.152, -0.032]
Safety Index	0.312	0.013	0.073	0.007*	[0.171, 0.415]
FDI Inflows (USD)	0.000	0.000	0.000	0.843	[-0.000, 0.000]
Note: Bootstrap results are based on 1,000 bootstrap samples. Significant at $p < 0.01$. Dependent Variable: Price-to-Income Ratio.					

[d.] Source: Own elaboration based on data integration from [23]-[25]

The above analysis provides a solid basis for the following conclusion: safety acts as a "nonfinancial capital" that is strongly associated with real estate value. From a scientific point of view, this means that the study offers new empirical evidence in the field of security for countries on the path of European integration, where institutional reforms directly affect the competitiveness and price structure of the national real estate market.

VI. Conclusion

The results obtained within the framework of the study confirm that the dynamics of housing affordability in the EU Member States and Georgia are a multifaceted process, where socio-institutional stability plays a crucial role. Based on the empirical results obtained from the model, we can confidently formulate the following fundamental conclusions:

1) Validation of the "safety premium": The study showed that the Safety index (X_1) is the most important structural driving factor affecting price formation ($B = 0.312$, $p = 0.007$). The results obtained confirm the theory that the real estate market considers institutional security to be the most important asset. This "Safety premium" is directly reflected in price growth, and this often exceeds the growth rate of household incomes and creates a barrier to housing affordability.

2) Macroeconomic paradigm shift: The statistical inertia of traditional economic variables such as GDP and foreign direct investment (FDI) suggests that the affordability crisis in modern transition markets is increasingly determined by institutional and social factors rather than purely monetary aggregates.

A. Recommendations for Policy Makers

Based on empirical evidence, we can formulate the following strategic recommendations for stabilizing the housing market:

1) Integrating institutional factors into urban planning: Urban development strategies should consider "Safety environment" as a nonfinancial instrument for market stabilization. It is necessary to strengthen public safety and increase institutional transparency in order to reduce the "artificial inflation" of real estate prices caused by local "safety premiums." If safety standards are uniform throughout the city, this will eliminate sharp price jumps in the so-called "safe zones."

2) Targeted housing strategies for transition economies: For countries like Georgia, which are embarking on the path of European integration, a balanced housing policy is vital. As institutional reforms improve safety ratings and, accordingly, increase property values, the state should implement social housing programs or rent control mechanisms. This avoids the forced displacement of low-income populations from "unsafe and expensive" urban centers.

3) Diversify affordability indicators: Regulators should move beyond traditional economic indicators and use socio-institutional indices (such as Safety Index and Quality of Life) when monitoring the housing market.

In summary, the scientific value of this study lies in the empirical confirmation of safety as an endogenous factor in price formation. By showing that institutional stability is a key driver of housing affordability, this study paves the way for future research to explore in more depth the impact of various institutional risks on the long-term sustainability of real estate markets.

B. Research Limitations and Future Directions

The primary limitation of this study arises from the unbalanced structure of the panel data and the reliance on subjective composite indicators, such as the Numbeo Safety Index, which reflect public perception rather than official institutional crime statistics. Additionally, due to the complete-case protocol constraints, the sample size is limited to 89 country-year observations. Future research should leverage high-frequency geospatial property listings and structural equation modeling (SEM) to unpack the underlying transmission mechanisms between institutional stability and localized housing markets.

References

[1] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, "Identifying real estate opportunities using machine learning," *Appl. Sci.*, vol. 8, no. 11, p. 2321, 2018. https://doi.org/10.3390/app8112321

[2] C. P. Hu, T. S. Hu, P. Fan, and H. P. Lin, "The urban blight costs in Taiwan: Evidence from the housing market," *Sustainability*, vol. 13, no. 1, p. 113, 2021. https://doi.org/10.3390/su13010113

[3] P. Taltavull de La Paz and M. White, "The sources of house price change: Identifying liquidity shocks to the housing market," *J. Eur. Real Estate Res.*, vol. 9, no. 1, pp. 98-120, 2016. https://doi.org/10.1108/JERER-11-2015-0041

[4] J. Salukvadze and O. Golubchikov, "City as a geopolitics: Tbilisi, Georgia - A globalizing metropolis in a turbulent region," *Cities*, vol. 52, pp. 39-54, 2016. https://doi.org/10.1016/j.cities.2015.11.013

[5] M. Renigier-Biłozor, A. Janowski, M. Walacik, and A. Chmielewska, "Human capital social index - real estate market analysis digitalization," *Land Use Policy*, vol. 113, p. 105876, 2022. https://doi.org/10.1016/j.landusepol.2021.105876

[6] S. Tripathi, "Macroeconomic determinants of housing prices: Evidence from cross-country panel data," MPRA Paper No. 98089, University Library of Munich, 2019.

[7] C. H. Yang, B. Lee, and Y. D. Lin, "Effect of money supply, population, and rent on real estate: A clustering analysis in Taiwan," *Mathematics*, vol. 10, no. 7, p. 1155, 2022. https://doi.org/10.3390/math10071155

[8] C. Leung, "Macroeconomics and housing: A review of the literature," *J. Hous. Econ.*, vol. 13, no. 4, pp. 249-267, 2004. https://doi.org/10.1016/j.jhe.2004.09.002

[9] A. Belke and J. Keil, "Fundamental determinants of real estate prices: A panel study of German cities," *Int. Adv. Econ. Res.*, vol. 24, no. 1, pp. 25-45, 2018. https://doi.org/10.1007/s11294-018-9671-2

[10] E. L. Glaeser and J. M. Nathanson, "Housing Bubbles," NBER Working Paper No. 20426, National Bureau of Economic Research, 2014. https://doi.org/10.3386/w20426

[11] N. Terterashvili and S. Shaburishvili, "A Study of the Factors Affecting the Real Estate Market with a Macroeconomic Context," *The Seybold Report*, vol. 18, no. 8, pp. 705-717, 2023. https://doi.org/10.5281/zenodo.8288943

[12] E. C. M. Hui and C. Y. Yiu, "Market dynamics of private residential real estate price – an empirical test in Hong Kong," *J. Financ. Manag. Prop. Constr.*, vol. 8, no. 3, pp. 155-165, 2003.

[13] Z. Adams and R. Füss, "Macroeconomic determinants of international housing markets," *J. Hous. Econ.*, vol. 19, no. 1, pp. 38-50, 2010. https://doi.org/10.1016/j.jhe.2009.10.005

[14] E. L. Glaeser, J. Gyourko, and R. E. Saks, "Why is Manhattan so expensive? Regulation and the rise in housing prices," *J. Law Econ.*, vol. 48, no. 2, pp. 331-369, 2005. https://doi.org/10.1086/429979

[15] K. Tsatsaronis and H. Zhu, "What drives housing price dynamics: cross-country evidence," *BIS Q. Rev.*, pp. 65-78, Mar. 2004.

[16] E. L. Glaeser and J. Gyourko, "The economic implications of housing supply," *J. Econ. Perspect.*, vol. 32, no. 1, pp. 3-30, 2018. https://doi.org/10.1257/jep.32.1.3

[17] Al-haimi, H. Khalid, N. H. Zakaria, and T. H. Jasimin, "Digital transformation in the real estate industry: A systematic literature review of current technologies, benefits, and challenges," *International Journal of Information Management Data Insights*, vol. 5, no. 1, p. 100340, 2025. https://doi.org/10.1016/j.jjimei.2025.100340

[18] H. Ahmed, T. A. Jilani, W. Haider, S. N. Hasany, M. A. Abbasi, and A. Masroor, "Producing Standard Rules for Smart Real Estate Property Buying Decisions based on Web Scraping Technology and Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 3, pp. 497-503, 2020. https://doi.org/10.14569/IJACSA.2020.0110363

[19] N. Naeem, I. A. Rana, and A. R. Nasir, "Digital real estate: a review of the technologies and tools transforming the industry and society," *Smart Construction and Sustainable Cities*, vol. 1, no. 1, p. 15, 2023. https://doi.org/10.1007/s44268-023-00016-0

[20] M. H. Hasan, M. A. Jahan, M. E. Ali, Y.-F. Li, and T. Sellis, "A Multi-Modal Deep Learning Based Approach for House Price Prediction," *arXiv preprint arXiv:2409.05335*, 2024. https://doi.org/10.48550/arXiv.2409.05335

[21] S. A. Rahman, N. H. Zulkifley, I. Ibrahim, and S. Mutalib, "Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 12, pp. 770-779, 2021. https://doi.org/10.14569/IJACSA.2021.0121291

[22] A. C. Cameron and D. L. Miller, "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, vol. 50, no. 2, pp. 317-372, 2015. https://doi.org/10.3368/jhr.50.2.317

[23] Numbeo, "Crime and Safety Index by Country," 2026. [Online]. Available: https://www.numbeo.com. [Accessed: Mar. 11, 2026].

[24] The World Bank, "World Development Indicators," 2026. [Online]. Available: https://data.worldbank.org. [Accessed: Mar. 11, 2026].

[25] Eurostat, "Housing price index and macroeconomic indicators," 2026. [Online]. Available: https://ec.europa.eu/eurostat. [Accessed: Mar. 11, 2026].

[26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994. https://doi.org/10.1201/9780429246593

[27] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2019. ISBN: 9780470526798

[28] R. Davidson and J. G. MacKinnon, "Bootstrap tests: How many bootstraps?" Econometric Reviews, vol. 19, no. 1, pp. 55-68, 2000. https://doi.org/10.1080/07474930008800459

[29] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 6th ed. Boston, MA: Cengage Learning, 2015. ISBN: 9781305270107

[30] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th international ed. New York, NY: McGraw-Hill, 2012. ISBN: 9780071330030

[31] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *J. Political Econ.*, vol. 82, no. 1, pp. 34-55, 1974. https://doi.org/10.1086/260169

An Enhanced Kalman Filter-Based Hybrid Battery Management System for Energy Optimization in IoT-Enabled Smart Agriculture

Mohd Kamir Yusof[1]*, Nur Yasmin Salleh[2], Senny Luckyardi[3],
Wan Mohd Amir Fazamin Wan Hamzah[4], Mustafa Man[5]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, 22200 Besut, Terengganu, Malaysia[1, 2, 4]
Universitas Komputer Indonesia, Jl. Dipatiukur No. 102-118, Kota Bandung, Jawa Barat 40132, Indonesia[3]
Faculty of Ocean Engineering Technology and Informatics,
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia[5]

***Abstract*—Battery-powered sensor nodes operating in a remote environment are used in IoT-based smart agriculture deployments, where the maintenance of these devices is impractical. Thus, precise SOC (state-of-charge) estimation is critical to enhance both energy efficiency and system stability. A new hybrid BMS for SOC estimation based on a transaction-oriented adaptive Kalman Filter that merges traditional voltage-based SOC estimation has been developed in this study. Results show that our proposed method removes noise sensitivity and reduces SOC drift subjected to dynamic IoT workloads by integrating model prediction with measurement correction. Application experiments based on 30 days' real sensor workload data indicate that the cumulative battery consumption index of the proposed algorithm decreased from 78% to 56%, and provided a performance improvement of approximately 22.7% compared to the baseline method. Moreover, the proposed method shows better stability in terms of daily consumption fluctuation and drift. The findings demonstrate that the improved Kalman Filter–based hybrid BMS serves as an efficient and effective alternative for achieving long-term energy optimization in IoT-enabled smart agriculture applications.**

Keywords*—Internet of Things; smart agriculture; battery management system; state-of-charge estimation; Kalman Filter; energy optimization

I. Introduction

Agriculture is the backbone industry for global food security and economic growth. A growing demand for food, coupled with climate change, has increased the need to develop more efficient and sustainable farming methods [1], [3]. To solve the aforementioned issues, smart agriculture leveraging Internet of Things (IoT) technologies has proven to be a promising solution that increases productivity via real-time monitoring, automated operations and data-driven decision making [2], [8], [11].

Climate change, environmental degradation, and increasing agricultural demand continue to affect global food sustainability and farming efficiency [4], [6], [7]. Furthermore, adoption barriers associated with cost, scalability, and infrastructure limitations remain significant challenges in smart agriculture deployment [5].

However, in spite of these remarkable benefits, IoT-based agriculture systems come up against important energy efficiency challenges. Most of the IoT devices are deployed in remote and open-field environments with access to power strings, thus have greater reliance on energy stored within batteries [10], [11]. Continuous sensing, on-device processing, and wireless communication drain the battery quickly, resulting in increased maintenance requirements and reduced system reliability [12].

Battery management systems (BMS) are crucial for monitoring battery health, ensuring safety during operation, and optimizing energy efficiency [13], [16]. One of their core functions, accurate estimation of the battery State of Charge (SOC), is crucial to optimizing energy management. Voltage-based methods and Coulomb counting are traditional SOC estimation approaches, which are highly sensitive to measurement noise, suffer from cumulative drift, and have low accuracy at dynamic working conditions [14], [15].

Consequently, KF-based SOC estimation methods have been widely utilized since they provide the advantages of combining system models and noisy sensor measurements [17], [18]. However, current KF implementations only focus on estimation accuracy and do not specifically consider optimization of battery consumption for IoT workloads applied to smart agriculture contexts. To fix this limitation, this study proposes an advanced Kalman Filter-based hybrid BMS to establish a direct relationship between SOC estimation accuracy and energy consumption optimization.

The key contributions of this work include:

1. A hybrid BMS architecture based on the voltage-based estimate and a Kalman filter correction approach.
2. An adaptive SOC modelling technique based on transactions for dynamic IoT workload variations.
3. Using 30-day real workload data for the experimental validation shows improved energy efficiency, stability, and less SOC drift.

Unlike existing SOC estimation approaches designed primarily for electric vehicle applications, the proposed method specifically targets low-power IoT smart agriculture workloads using transaction-based adaptive covariance adjustment to improve long-term energy optimization and SOC stability.

*Corresponding author

The remainder of this study is organized as follows: Section II reviews related work on IoT energy challenges and SOC estimation techniques. Section III presents the proposed hybrid battery management system architecture and transaction-based adaptive Kalman Filter methodology. Section IV describes the experimental setup, hardware configuration, and data collection process. Section V discusses the experimental results, comparative analysis, and system performance evaluation. Section VI details the limitations of the study. Finally, Section VII concludes the study and outlines future research directions.

II. Related Work

A. IoT Energy Challenges in Smart Agriculture

The swift adoption of Internet of Things (IoT), also known as smart agriculture technologies, has led to its myriad applications, such as automated irrigation, crop & soil health monitoring, and environmental sensing. Typically, these systems do operate on distributed sensor nodes placed around an open-field environment, a greenhouse, or a remote farm. While the productivity and resource utilization rise with IoT-based agriculture, energy efficiency is still one of the major critical problems due to the limited power available for sensor nodes [11], [12]. Power is a major concern in this domain as most of the agricultural IoT deployments are battery-powered, which makes it either difficult or costly to replace batteries frequently [10]. Energy consumption mainly comes from continuous sensing, local processing, and wireless data transmission. This is especially true, since communication overhead often accounts for a considerable portion (more than 90% in some cases) of the energy cost [8] associated with such workloads. Outdoor environments (i.e., temperature variability, humidity, and erratic connectivity) result in quick battery drain of these systems as well as low overall reliability of the system. Energy in IoT-based smart agriculture systems is desired to be sustained by the long-term existence, hence energy-aware level monitoring and power optimization methodologies are required.

B. SOC Estimation Techniques

BMS is able to enhance its performance, but also control and ensure battery and overcharging in an exceptional way. In order to optimize the energy, BMS will calculate the state of charge (SOC) of the battery pack for every essential task and used for remaining battery life estimation [13], [16]. Meanwhile, SOC estimation algorithms can also be grouped into measurement mode, model-based products, data-driven techniques, and filter-based mechanisms [14], [15]. By this combination of groups, voltage-based SOC estimation becomes simple and low in computation. It is greatly affected by loading conditions, aging of the battery, and measuring noise [14]. Earlier studies on battery SOC determination have also highlighted the limitations of conventional voltage-based approaches and emphasized the need for more robust estimation techniques [20]. At the same time, SOC can also be estimated by Coulomb counting to track the previous flow over time. However, a periodic calibration based on cumulative error is needed to avoid long-term drift [15]. Therefore, some of the function and demonstration using SOC estimation techniques still have a limitation when transferring to operational conditions. That way, SOC are considered active areas of research for improvement especially for IoT applications with limited computational resources and energy budgets [9], [10].

C. Kalman Filter-Based SOC Estimation

The Kalman Filter (KF) approach is one of the most theoretically established methods for SOC estimation because it combines system modelling with noisy sensor measurements [16], [17]. Early work by Plett demonstrated the effectiveness of Extended Kalman Filtering (EKF) for battery management systems and lithium-ion battery SOC estimation [21]. To account for battery model nonlinearities, EKF and Unscented Kalman Filter (UKF) variants have subsequently been proposed [19], [21], [23]. Furthermore, comparative studies have shown that adaptive Kalman filtering techniques can outperform conventional SOC estimation approaches under varying operating conditions by dynamically adjusting noise covariance parameters [22]. Adaptive KF methods, therefore, improve robustness against measurement uncertainty and model mismatch while reducing SOC drift and transient voltage fluctuation effects [17], [22].

D. Research Gap and Motivation

Demonstrated to be very accurate and robust, the KF-based SOC estimation approaches are nowadays more applied to obtaining a (possible) more accurate estimation for the SOC than being able to explicitly extract an energy optimization link between the SOC estimation, and performing these optimizations actually on location and paradigm that targets IoT workloads of sensors. For example, in smart agriculture realized through IoT, sensor nodes have different energy consumption patterns owing to periodic sensing, burst transmission of sensed data from the field to data centers or cloud servers for aggregation and processing, sleep scheduling, and sudden environmental conditions [3]. If traditional estimation methods are used, these time-variant workloads might cause SOC drift, inaccurate estimation, and unsuitable use of battery power. Furthermore, research is limited regarding hybrid SOC estimation techniques that combine voltage-based methods with adaptive filtering in smart agriculture scenarios. For this, a low-cost and energy-aware SOC estimation framework can be proposed to avoid drift and ensure stability for transaction-driven IoT workloads. This interaction causes the proposed hybrid BMS architecture, combining a voltage-based estimation strategy with transaction-based adaptive Kalman Filter (AKF), to provide embedded smart agriculture deployments a way to learn long-term energy optimization.

III. System Architecture and Methodology

A. Overall System Architecture

The purpose of battery management system (BMS) is to manage the battery used in IoT-based Smart Agriculture sensor nodes functioning under limited capacity. We present a system consisting of a lithium-ion battery pack, sensing and measurement unit, SOC estimation module, and data processing and monitoring layer. The measurement subsystem constantly samples battery parameters, including voltage, current and temperature, to estimate SOC in a time continuous way.

Fig. 1 illustrates the baseline voltage-based SOC estimation architecture used in the conventional monitoring framework. Although such method is computation inexpensive and can be

applied to low-power IoT devices, it is susceptible to noisy voltage operation and load variations. Meanwhile, Fig. 2 presents the proposed hybrid adaptive Kalman Filter architecture, which integrates voltage-based estimation with transaction-aware adaptive filtering. The SOC estimation module predicts and updates based on the battery model and real-time measurements in this architecture.

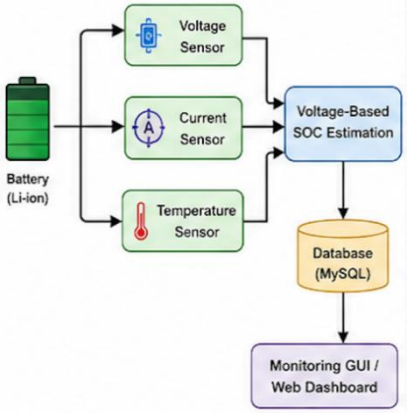


Fig. 1. Baseline voltage-based SOC estimation architecture.

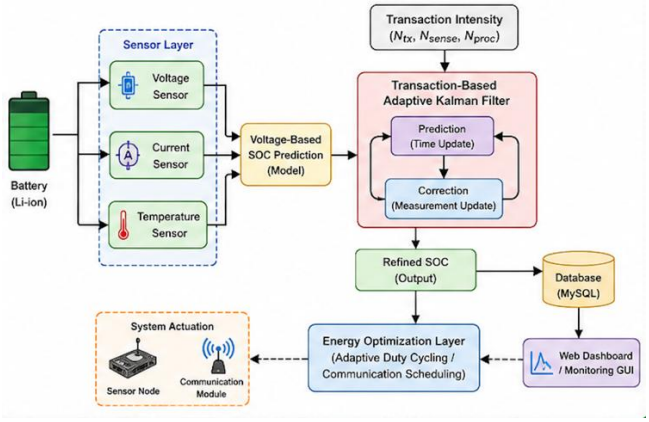


Fig. 2. Proposed hybrid adaptive Kalman Filter architecture.

In order to verify the performance of the proposed hybrid BMS, two estimation algorithms are executed and compared with each other in parallel:

1) Algorithm A: Charge status (SOC) estimation based on baseline voltage.

2) Algorithm B: Moving-based transaction adaptive Kalman Filter hybrid SOC estimation.

B. Battery SOC State-Space Model

SOC estimation is modelled as a state-space model in the implementation of the Kalman Filter. The SOC dynamic equation can be written as:

$$SOC_k = SOC_(k\text{-}1) - (I_k \Delta t / C) = w_k \quad (1)$$

where, SOC_k is the SOC at time step k, I_k is the measured current, Δt is the sampling interval, C is the rated battery capacity, and w_k represents process noise.

The voltage measurement model is expressed as:

$$V_k = f(SOC_k) + v_k \quad (2)$$

where, V_k is the measured terminal voltage, f(·) represents the battery voltage model, and v_k denotes measurement noise.

C. Baseline SOC Estimation (Algorithm A)

Algorithm A or Algorithm 1 follows a conventional voltage-based SOC estimation approach with Coulomb counting integration. This technique uses the measured voltage, current, and temperature values to estimate the SOC of the battery. So a voltage compensation factor is introduced to represent measurement error under the open-circuit and closed circuit conditions, and a temperature compensation coefficient is kept for the SOC value in terms of its influence by the change of temperature.

Algorithm 1: Voltage-Based SOC Estimation (Algorithm A)

Input: Measured voltage V(t), current I(t), temperature T(t), battery capacity C, internal resistance R_int, initial SOC SOC_init, temperature coefficient α, reference temperature T_ref, voltage correction factor ΔV, sampling interval Δt.

Output: Estimated SOC SOC(t).

Step 1 (Initialization):
SOC(0) = SOC_init
Step 2 (Voltage Correction):
V_corr(t) = V(t) − ΔV
Step 3 (Charge Integration):
Q(t) = ∫_0^t I(τ)d τ
Discrete form: Q(t) = Σ I(k)Δt
Step 4 (SOC Estimation):
SOC(t) = SOC_init − Q(t)/C
Step 5 (Temperature Compensation):
SOC_temp(t) = SOC(t)[1 + α(T(t) − T_ref)]

Step 6 (Multi-Cell Voltage Averaging):
V_avg = (1/n) Σ x_i
Return: SOC_temp(t)

D. Proposed Transaction-Based Adaptive Kalman Filter Estimation (Algorithm B)

Algorithm B or Algorithm 2 estimates SOC using two transaction-based workload models. Additionally, for IoT workloads, it uses the Kalman Filter correction mechanism to minimize SOC drift. Rather than being continuous, as in sensing and data transmission, discrete operational events and activity periods determine the battery consumption of IoT smart agriculture systems. These activities, which are represented as energy transactions, cause battery consumption. In this context, “transaction" is the term used to describe a unitary energy event associated with IoT activities, such as sensing, processing, and wireless transmission. The noise covariance of the Kalman process is adaptively adjusted based on transaction intensity, since transactions include multiple sensing cycles and communication events per unit of time.

Algorithm 2: Transaction-Based Adaptive Kalman Filter SOC Estimation (Algorithm B)

Input: V(t), I(t), T(t), C, SOC_init, ΔV, α, T_ref, Δt, process noise covariance Q, measurement noise covariance R.

Output: Kalman-filtered SOC estimate SOC(t).

Step 1 (Initialization):

$\hat{x}_0 = SOC_{init}, P_0 = P_{init}$

Step 2 (Voltage Correction):

$V_{corr}(t) = V(t) - \Delta V$

Step 3 (SOC Prediction):

$SOC_{pred}(t) = SOC_{init} - (\int_0^t I(\tau)d\tau)/C)$

Step 4 (Temperature Compensation):

$SOC_{temp}(t) = SOC_{pred}(t)[1 + \alpha(T(t) - T_{ref})]$

Step 5 (Kalman Prediction):

$\hat{x}_{k|k-1} = SOC_{temp}(t)$

$P_{k|k-1} = P_{k-1|k-1} + Q$

Step 6 (Kalman Update):

$y_k = V_{corr}(t) - V_{model}(\hat{x}_{k|k-1})$

$K_k = P_{k|k-1}/(P_{k|k-1} + R)$

$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k y_k$

$P_{k|k} = (1 - K_k)P_{k|k-1}$

Step 7 (Pack SOC Output):

$SOC(t) = (1/N) \Sigma SOC_{cell_i}(t)$

Return: SOC(t)

E. Transaction-Based Adaptive Noise Adjustment

To improve SOC estimation robustness under dynamic IoT workloads, the proposed Adaptive Kalman Filter (AKF) modifies the process noise covariance Q according to transaction intensity. In IoT-enabled smart agriculture, battery consumption varies depending on sensing frequency, wireless communication activity, local processing tasks, and environmental events. Therefore, the uncertainty of the SOC prediction model changes over time.

The adaptive process noise covariance is defined as:

$$Q_k = Q_0 (1 + \beta T_k)$$

where,

Q_k represents the adaptive process noise covariance at time step k,

Q_0 is the nominal process noise covariance,

β is the workload sensitivity coefficient,

T_k denotes the normalized transaction intensity.

The transaction intensity T_k is computed as:

$$T_k = (N_{tx} + N_{sense} + N_{proc})/N_{max}$$

where,

N_{tx} is the number of wireless transmission events,

N_{sense} is the number of sensing operations,

N_{proc} is the number of local processing tasks,

N_{max} is the maximum expected transaction count during one sampling interval.

When transaction activity increases, the value of Q_k becomes larger, allowing the Kalman Filter to react faster to rapid SOC changes. During low activity periods, Q_k decreases, stabilizing the SOC estimate and reducing unnecessary fluctuations.

This adaptive mechanism enables the proposed hybrid BMS to maintain estimation stability while remaining responsive to highly dynamic IoT sensor workloads commonly observed in smart agriculture deployments.

F. Hybrid SOC Output and Refinement

In such a hybrid mechanism, the final SOC is obtained from a Kalman Filter output that has been filtered. The use of a voltage-based methodology allows for an initial estimation, and a Kalman filter adjusts this prediction error while also reducing noise. Enhanced SOC estimate, drift mitigation, and stability of the proposed hybrid mechanism provide significant contributions towards long-run IoT-based smart agriculture procedures.

IV. Experimental Setup

A. Hardware and Software Configuration

The database used in this experiment is MySQL. It was chosen for its reliability, ease of use, and ability to query and manipulate large amounts of data. Furthermore, it can quickly organize and analyze large amounts of information. Meanwhile, an Arduino R4 device programmed in C can control a variety of sensors and components. For the hybrid BMS model experiment, this combination simplifies the integration and control of the hardware and database.

Table I summarizes the hardware and software configuration used in the experimental implementation of the proposed hybrid BMS framework.

TABLE I. Hardware and Software Configuration

Parameter	Specification
Microcontroller	Arduino UNO R4 WiFi
Battery Type	3.7 Lithium-ion Battery
Battery Capacity	2200 mAh
Voltage Sensor	INA219 Voltage Sensor Module
Current Sensor	INA219 Current Monitoring Sensor
Temperature Sensor	DS18B20 Digital Temperature Sensor
Communication Module	ESP32 WiFi Module
Programming Language	Embedded C (Arduino IDE)
Database	MySQL
Data Storage Platform	Ubuntu Server 22.04
Experimental Duration	30 days
Sampling Interval	1 hour
Operating Voltage	3.3 V – 5 V
Wireless Connectivity	WiFi IEEE 802.11 b/g/n

B. Data Collection and Processing

Table II presents sample hourly SOC data collected during the experimental evaluation over the 30-day monitoring period.

Table III presents the sample hourly battery usage computation derived from the SOC measurements, while

Table IV summarizes the daily battery usage recorded throughout the 30-day experimental period.

TABLE II. SAMPLE HOURLY SOC DATA COLLECTED DURING EXPERIMENT.

Day	Hour	SOC_A (%)	SOC_B (%)
1	00	100.00	100
1	01	98.8	99.1
1	02	97.2	98.4
…	…	…	…

TABLE III. SAMPLE HOURLY BATTERY USAGE COMPUTATION

Hour	Used A (%)	Used B (%)
1	1.4	0.9
2	1.4	0.7
3	1.6	1.0

TABLE IV. DAILY BATTERY USAGE OVER 30-DAY EXPERIMENT

Day	Algorithm A Daily Used (%)	Algorithm B Daily Used (%)
1	32.4	24.8
2	30.9	25.1
3	34.2	26.0
…	…	…

C. Performance Metrics

Table V summarizes the overall battery consumption and average daily energy usage achieved by Algorithm A and Algorithm B, enabling a direct comparison of their energy optimization performance.

Table VI presents the stability and drift characteristics of both SOC estimation methods, including standard deviation and drift spike measurements, to evaluate long-term estimation reliability.

Table VII reports the SOC estimation accuracy in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), providing a quantitative assessment of the effectiveness of the proposed hybrid Kalman Filter approach.

TABLE V. OVERALL BATTERY CONSUMPTION AND AVERAGE DAILY ENERGY USAGE.

Algorithm	Total Battery Consumption (%)	Average Daily Consumption B (%)
Algorithm A	78	32.6
Algorithm B	56	25.2

TABLE VI. STABILITY AND DRIFT COMPARISON

Metric	Algorithm A	Algorithm B
Standard Deviation (%)	3.8	1.6
Drift Spike	41.2	29.5
Drift Trend	High	Low

TABLE VII. SOC ESTIMATION ERROR COMPARISON

Metric	Algorithm A	Algorithm B
MAE (%)	5.6	2.3
RMSE (%)	7.9	3.4

V. RESULTS AND DISCUSSION

A. Battery Management System Implementation

In order to test the introduced hybrid SOC estimation framework, a prototype battery management system (BMS) was created. The proposed system monitors the parameters of the battery, such as voltage, current, and temperature, in real-time, to store measured data in a database for future analysis. Moreover, a GUI (Graphical User Interface) was developed to display SOC estimation results, battery consumption trends, and sensor readings. Fig. 3 compares daily battery consumption between Algorithm A and Algorithm B over the 30-day experiment, and Fig. 4 illustrates the drift spike comparison between the estimation methods.

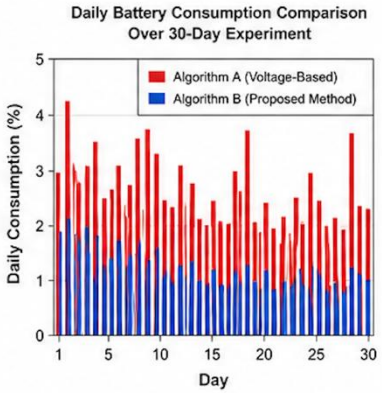


Fig. 3. Daily battery consumption comparison over a 30-day experiment.

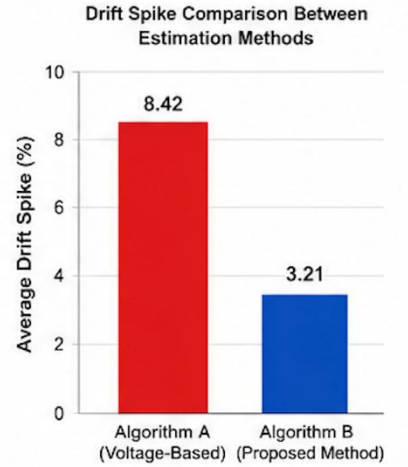


Fig. 4. Drift spike comparison between estimation methods.

B. Definition of Battery Consumption Metric

The battery consumption percentage is defined as the SOC reduction over time and is used to evaluate the performance of energy usage. This study does not use “Battery % Used” as the percentage of battery's SOC value but rather as the difference

between SOC estimates for certain cycles. This hourly battery consumption is calculated as:

$$Used_hour(k) = SOC_(k\text{-}1) - SOC(k) \quad (3)$$

Similarly, daily battery usage is computed as the accumulated hourly SOC reduction over a 24-hour period:

$$Used_day(d) = \Sigma\, Used_day(d), d = 1\ to\ 30 \quad (4)$$

This definition enables consistent comparison between Algorithm A and Algorithm B in terms of discharge behavior and SOC stability.

C. Comparison of Battery Consumption Performance

The comparison between Algorithm A (baseline voltage-based SOC estimation) and Algorithm B (transaction-based adaptive Kalman Filter hybrid SOC estimation) was performed based on workload data collected over 30 days. The experimental results showed that Algorithm A produced higher battery consumption compared to Algorithm B. Algorithm A showed a total battery usage of 78%, while Algorithm B exhibited an aggregate battery usage of 56%. This shows that the estimated battery usage is reduced by roughly 22.7% for the proposed method compared to the baseline approach.

Moreover, the final average daily power consumption values for Algorithm A and Algorithm B are 32.6% and 25.2%, respectively. These results show that the hybrid method proposed offers more energy-efficient SOC estimation, a parameter relevant to prolonging sensor node life in IoT smart agriculture implementations.

D. Stability and Drift Analysis

Besides lowering overall battery consumption, Algorithm B also displayed enhanced stability in SOC estimation. Standard deviation was used to measure the variability of daily battery usage. The higher value of 3.8% was the standard deviation for Algorithm A, whereas Algorithm B had a lower value of 1.6%. This means that the proposed Kalman Filter refinement mechanism is more stable during dynamic IoT workloads.

The proposed method proved robust for models with strongly drifted classes. Both algorithms showed a drift spike of 41.2% in Algorithm A, and decreased to 29.5% in Algorithm B. This enhancement is due to the working mechanism of Kalman correction, which mitigates errors and iteratively updates the SOC prediction as more measurements become available. The lower drift behavior plays an important role in long-term deployments since drift accumulation can give rise to inaccurate battery depletion prediction and inefficient energy scheduling.

E. Estimation Accuracy Evaluation

SOC estimation results were then compared to reference discharge measurements to assess the accuracy of the estimations. Evaluation metrics, including mean absolute error (MAE) and root mean square error (RMSE), were used. Algorithm A achieved a MAE of 5.6% while Algorithm B's MAE was significantly lower at 2.3%. The RMSE also showed improvement, going from 7.9% in Algorithm A to 3.4% in Algorithm B.

The presented transaction-based adaptive Kalman Filter method effectively improves SOC estimation accuracy through cumulative drift minimization and noise sensitivity reduction, as evidenced by these cumulative results. The enhanced precision additionally aids battery consumption prediction, leading to energy optimization within IoT sensor nodes.

F. Comparison with Existing KF-Based SOC Estimation Methods

To further evaluate the effectiveness of the proposed transaction-based adaptive Kalman Filter method, its performance was compared with several SOC estimation approaches reported in previous studies, including conventional Kalman Filter (KF), Extended Kalman Filter (EKF), and Adaptive Kalman Filter techniques.

The selected comparison methods are based on established SOC estimation approaches reported in the literature, including EKF-based battery management systems [21], adaptive Kalman filtering techniques [22], and adaptive unscented Kalman filtering methods [23].

Table VIII compares the proposed method with existing SOC estimation techniques in terms of estimation accuracy, computational complexity, and suitability for IoT-based smart agriculture applications.

TABLE VIII. SOC ESTIMATION ERROR COMPARISON

Method	MAE (%)	RMSE (%)	Complexity	IoT Suitability
Voltage-Based	5.6	7.9	Low	Moderate
Conventional KF	3.8	5.2	Moderate	Moderate
EKF	3.1	4.5	High	Low
Adaptive KF	2.7	3.9	Moderate	Good
Proposed Method	2.3	3.4	Low-moderate	Excellent

The comparison demonstrates that the proposed method achieved the lowest MAE and RMSE values while maintaining relatively low computational complexity suitable for embedded IoT sensor nodes. Unlike EKF-based approaches that require higher nonlinear computation, the proposed transaction-based adaptive Kalman Filter provides improved SOC stability and energy optimization with a lightweight implementation suitable for smart agriculture deployments.

G. Discussion

In general, the experimental performance shows that Algorithm B outperforms Algorithm A in energy efficiency, being more stable and with an improved estimation accuracy. The decrease in overall battery consumption suggests that more realistic state-of-charge (SOC) tracking can be achieved with the proposed method, thus preventing unnecessary overestimation of the battery's depletion. It can be observed from the enhanced stability and decreased drift spikes that the adaptive Kalman Filter algorithm in this study is helpful for reducing SOC fluctuation under dynamic load condition.

The proposed hybrid BMS is a suitable and cost-effective approach to long-term monitoring in sensor nodes powered by energy limited resources present in IoT-enabled smart agriculture systems. This ensures greater adaptability to fluctuation in sensing and transmission activity, making

effective utilization of charge within battery limit and enhancing longevity and maintenance period.

H. Practical Implications for Smart Agriculture

In general, the experimental performance shows that the proposed hybrid BMS architecture provides several practical benefits for IoT-enabled smart agriculture systems. Since many agricultural monitoring systems are deployed in remote farms and open-field environments, minimizing battery replacement frequency is essential for reducing maintenance costs and improving operational reliability.

The proposed transaction-based adaptive Kalman Filter enables more accurate prediction of remaining battery energy, allowing sensor nodes to optimize sensing intervals and communication schedules dynamically. This capability is particularly beneficial for precision agriculture applications involving soil moisture monitoring, irrigation automation, greenhouse environmental sensing, and crop health monitoring.

Furthermore, improved SOC stability can reduce unexpected sensor downtime caused by inaccurate battery depletion prediction. The lightweight computational requirement also makes the proposed method suitable for low-cost embedded microcontrollers commonly used in agricultural IoT deployments.

VI. Limitations of the Study

Although the proposed hybrid BMS demonstrates improved SOC estimation performance, several limitations remain. First, the experimental evaluation was conducted using a single battery configuration and a limited number of IoT sensor nodes. Different battery chemistries and larger-scale deployments may exhibit different performance characteristics.

Second, the current implementation focuses on WiFi-based communication workloads. Other communication technologies, such as LoRaWAN, ZigBee, and NB-IoT, may introduce different energy consumption patterns that require further investigation.

Finally, the proposed adaptive Kalman Filter relies on manually selected tuning parameters, including the workload sensitivity coefficient β. Future research should investigate self-optimizing parameter tuning methods using machine learning or reinforcement learning approaches.

VII. Conclusion and Future Work

This study proposed an enhanced Kalman Filter-based hybrid battery management system (BMS) for SOC estimation and energy optimization in IoT-enabled smart agriculture applications. The proposed method combines voltage-based SOC estimation with a transaction-based adaptive Kalman Filter correction mechanism to improve estimation stability and reduce SOC drift under dynamic IoT workloads.

Experimental evaluation using real sensor workload data collected over a 30-day period demonstrated that the proposed method reduced cumulative battery consumption from 78% to 56%, corresponding to approximately 22.7% improvement compared to the baseline voltage-based approach. In addition, the proposed method achieved lower MAE and RMSE values, improved SOC stability, and reduced drift spikes.

The adaptive transaction-based noise covariance adjustment enabled the Kalman Filter to dynamically respond to changing sensing and communication activities, making the proposed approach particularly suitable for energy-constrained IoT sensor nodes deployed in remote agricultural environments.

Future work will focus on extending the proposed framework to multi-node IoT deployments, integrating LoRaWAN-based communication systems, and comparing the proposed method against EKF, UKF, and machine learning-based SOC estimation techniques. Additional investigation into self-adaptive parameter optimization and renewable energy integration is also planned.

Acknowledgment

The authors would like to express their sincere gratitude to the Fundamental Research Grant Scheme (FRGS), the Ministry of Higher Education (Kementerian Pendidikan Tinggi, KPT), Malaysia, for providing financial support for this research under Grant No. FRGS/1/2022/ICT02/UNISZA/02/6.

References

[1] K. Kour and S. Arora, “Recent developments of the Internet of Things in agriculture: A survey,” IEEE Access, vol. 10, pp. 437–456, 2022.

[2] J. Chen and S. Yang, “A review of smart farming: IoT-based agriculture,” Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 8, pp. 3011–3032, 2019.

[3] P. Malhi, R. Kaur, and A. Singh, “Population growth and food demand: Challenges for sustainable agriculture,” Sustainable Computing: Informatics and Systems, vol. 30, pp. 100–109, 2021.

[4] M. Ahmad and K. Habib-ur-Rahman, “Impact of climate change on agricultural productivity: A global perspective,” Environmental Challenges, vol. 7, pp. 100–118, 2022.

[5] R. Patel, S. Shah, and P. Mehta, “Adoption barriers of smart agriculture technologies,” Computers and Electronics in Agriculture, vol. 170, pp. 105–113, 2020.

[6] M. Girip, A. Rahman, and S. Islam, “Environmental impacts of conventional agriculture practices,” Journal of Cleaner Production, vol. 252, pp. 119–128, 2020.

[7] M. Zaib, A. Khan, and H. Farooq, “Agricultural stress factors under climate variability,” Sustainability, vol. 15, no. 3, pp. 212–223, 2023.

[8] M. Maddikunta et al., “Internet of Things: A survey on enabling technologies, protocols, and applications,” IEEE Access, vol. 8, pp. 178–198, 2020.

[9] S. Shah, A. Kanhere, and R. Jadhav, “IoT-based monitoring systems for agriculture,” Procedia Computer Science, vol. 170, pp. 971–978, 2019.

[10] H. Tekinerdogan, “Farm Management Information Systems: Architecture and challenges,” Information Processing in Agriculture, vol. 6, no. 2, pp. 235–247, 2019.

[11] A. Tzounis, N. Katsoulas, T. Bartzanas, and C. Kittas, “Internet of Things in agriculture, recent advances and future challenges,” Biosystems Engineering, vol. 164, pp. 31–48, 2017.

[12] S. Deshpande, M. Sharma, and P. Rane, “IoT applications in livestock monitoring: A survey,” Computers and Electronics in Agriculture, vol. 181, pp. 105–113, 2021.

[13] J. González, M. Lopez, and R. Perez, “Battery energy storage systems for clean energy applications,” Renewable and Sustainable Energy Reviews, vol. 154, pp. 111–121, 2022.

[14] Y. Zhang, J. Xiong, and H. Wang, “State of charge estimation of lithium-ion batteries: A review,” Journal of Power Sources, vol. 383, pp. 92–104, 2018.

[15] Y. Meng, S. Li, and Z. Chen, “SOC estimation methods for lithium-ion batteries,” IEEE Transactions on Industrial Electronics, vol. 65, no. 7, pp. 5479–5490, 2018.

[16] M. H. Lipu, M. Hannan, A. Hussain, and A. Mohamed, "Battery management system for electric vehicles: A review," Renewable and Sustainable Energy Reviews, vol. 146, pp. 111–132, 2021.

[17] J. Yu, H. He, and X. Xiong, "Joint SOC and SOH estimation using Kalman filtering," IEEE Transactions on Power Electronics, vol. 34, no. 4, pp. 3250–3261, 2019.

[18] R. Xu, Z. Chen, and B. Li, "Kalman filtering techniques for battery SOC estimation," Journal of Power Sources, vol. 196, no. 10, pp. 4719–4728, 2012.

[19] Y. Peng, L. Wang, and M. Chen, "Advanced Kalman filter-based battery management systems," IEEE Access, vol. 10, pp. 55123–55135, 2022.

[20] S. Piller, M. Perrin, and A. Jossen, "Methods for state-of-charge determination and their applications," Journal of Power Sources, vol. 96, no. 1, pp. 113–120, 2001.

[21] G. L. Plett, "Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs," Journal of Power Sources, vol. 134, no. 2, pp. 252–261, 2004.

[22] X. Hu, F. Sun, and Y. Zou, "Comparison between adaptive Kalman filters and conventional methods for battery SOC estimation," Energy Conversion and Management, vol. 59, pp. 31–39, 2012.

[23] M. Partovibakhsh and G. Liu, "An adaptive unscented Kalman filtering approach for online estimation of battery state of charge," IEEE Transactions on Industrial Electronics, vol. 62, no. 9, pp. 5903–5911, 2015.

Enhanced Real-Time Fire Detection Systems Using Deep Learning and Differentiating Between Dangerous and Non-Dangerous Fires

Mohamed Youssef[1], Mohamed Marie[2], Sarah Naiem[3]
Department of Computer Science, Faculty of Computer Science and Artificial Intelligence, Capital University, Cairo, Egypt

***Abstract*—Fire is a major hazard in many disasters, creating risks to public safety and the surrounding environment. This study aims to improve the accuracy and reliability of fire detection using deep learning, addressing the key limitations of traditional sensor systems, such as latency and poor adaptability. The proposed model presents a customized Fire-Smoke-YOLOv8x Model trained from scratch on a dataset of 100,000 images representing diverse fire and smoke conditions paired with adaptive algorithms to differentiate dangerous from non-dangerous fire scenarios and achieved a performance with 98.2 % precision, 97.3% recall, and 94% mAP@50. The model processes RGB video under varied lighting and environments and achieves strong detection results across a wide range of fire and smoke scenarios. Incorporating thermal or multispectral data, as noted in future work, could further improve performance under extremely low visibility conditions. This framework supports real-time surveillance in smart cities, transport hubs, and industrial safety, where fast and accurate detection is critical. It combines fine-grained detection with risk-aware post-processing, providing a high-performance and scalable solution for real-world fire detection.**

***Keywords*—Deep learning; fire detection; smoke detection; computer vision; real-time detection; attention mechanism; fire risk classification**

I. Introduction

Human history has been shaped by a range of natural and human-caused disasters that threaten lives and disrupt communities, often resulting from environmental shifts that have led to widespread loss and destruction. Among these events, fire is especially dangerous due to its rapid spread and unpredictable behavior [1],[2],[3].

Fires are one of the most critical threats to public safety and community growth. They present risks to both cities and rural areas, pose a danger to human life and property, causing fear among people, damaging homes and buildings, destroying farms and forests, spreading quickly in dry conditions, harming animals, reducing air quality, and making recovery slow and difficult for affected communities [4], [5], [6].

Exacerbating this crisis, rising temperatures and long periods without rain make it easier for wildfires to start. Additional things, such as rising sea surface temperatures and extreme weather events, have also been linked to increased fire activity across urban and natural ecosystems [7]. Because of all that, strengthening early fire detection capabilities is becoming critical to emergency response strategies. Efforts have focused on deploying advanced, real-time fire detection systems that provide faster alerts, support early detection, and improve overall emergency response efficiency [8].

Factors such as global warming and land-use changes have contributed to the increased frequency and intensity of fire-related events. Recent years have shown a rise in extreme weather conditions, including heatwaves, which create favorable environments for fire ignition and spread. Many regions have reported unusual and large-scale fires, highlighting the growing need for reliable early detection systems. Early fire identification is critical because rapid detection enables timely intervention, reduces potential damage, and helps protect both lives and property. Effective early warning systems can support firefighters by providing prompt alerts in high-risk areas, by this way limiting the spread of fires before they escalate [9], [10]. To reduce the risk of loss of life and damage to property, it is important to detect fires as early as possible, deploy warning systems in at-risk locations, and intervene quickly to limit fire spread [11],[12]. Conventional fire detection technologies such as heat sensors, smoke sensors, and threshold-based alarm systems have been widely used for decades. However, their performance can be limited under varying environmental conditions, resulting in delayed or unreliable detection [13].

Recent advancements in computer vision provide promising alternatives. Deep Learning Models can analyze complex visual patterns, allowing better detection of flames and smoke under diverse lighting and environmental conditions [14],[15],[16].

II. Related Works

A wide range of approaches has been proposed for fire and smoke. Traditional fire detection methods typically rely on sensors that are activated only after the fire reaches a certain stage, meaning that detection often occurs after the fire has already spread significantly within the environment [17],[18]. This limitation has motivated researchers to shift toward intelligent vision-based systems, where models are trained to visually detect fire and smoke at earlier stages before widespread propagation occurs [19],[20]. Early detection is critical in reducing damage and improving response time, especially in scenarios where fires may be difficult to recognize in their initial phases [21]. Also, while sensor-based systems are commonly deployed in enclosed environments, their applicability in open or outdoor settings is limited.

This creates a need for intelligent detection systems capable of monitoring such environments using visual data, enabling timely identification and response [22],[23].

To address these limitations, several advanced fire and smoke detection methods have been introduced, as summarized in Table I

TABLE I. SUMMARY OF REPRESENTATIVE PRIOR FIRE AND SMOKE DETECTION STUDIES

	Authors names	Approach Category	Core Technique(s) and Features Used	Reported Strengths	Limitations
1	Chettoui, M., & Akhloufi, M. A. [24]	Deep learning–based fire/smoke detection using YOLOv7/YOLOv8 with transfer learning.	on a >11k image fire–smoke dataset; evaluation at IoU=0.5; trained using transfer learning with optimized hyperparameters	Best model is (YOLOv8x): mAP@0.5 = 92.6, Precision = 95.4, Recall = 84.8.	Evaluation is performed on a still image dataset; no video-based evaluation.
2	Prakash M., Neelakandan, S., Tamilselvi, M., Velmurugan, S., Priya, S. B., & Martinson, E. O. [25]	Deep learning (transfer learning)	Inception-v3 transfer learning to classify satellite/aerial images into fire vs. non-fire	Accuracy = 97.55%, Precision = 94.19%, Recall = 96.44% and F1 = 93.33% on their dataset.	No localization capability Trained and evaluated exclusively on satellite/aerial imagery, limiting generalization to indoor or ground-level scenes.
3	Vignesh R. P., Darshan Kumar P., Rajasekar S., & Prakash K. [26]	Fire and smoke detection using YOLO-based object detection	Real-time fire and smoke detection using YOLOv8 on annotated image datasets	Highlights YOLOv8's suitability for practical deployments (industrial/urban/forest). Demonstrates good accuracy on their custom dataset.	tested only on static images (no video/temporal robustness); limited dataset diversity.
5	S. Rahman, S. M. H. Jamee, J. K. Rafi, J. S. Juthi, S. A. Aziz and J. Uddin [27].	Deep learning (object detection; YOLOv8)	Fine-tune YOLOv8 on a 5,700-image custom dataset (fire/smoke / fire+smoke)	Real-time detection with 68.3% precision, 54.6% recall, and 60.7% F1; Outperformed several conventional CNN baselines.	Small dataset (~5.7k images); limited generalization; modest recall; no segmentation capability; performance drops in complex smoke conditions.
6	Mukhiddinov, M., Abdusalomov, A. B., & Cho, J. [28].	Deep learning (object detection; optimized YOLOv5m)	YOLOv5m enhanced with anchor-box clustering (K-means++), SPPF-plus, feature pyramid enhancement, pruning, and transfer learning;	Achieves AP@50 = 81.5%, AP = 73.6% on UAV-based wildfire smoke	UAV-only wildfire dataset (~6000 frames); limited generalization to indoor/ground fires; no FPS reported; higher model complexity.
7	N. Borges, L. Fonseca, P. S. Barreto, E. A. P. Alchieri, M. F. Caetano, P. A. Resende, L. Brandão and L. Vieira. [29].	Two-phase deep-learning smoke detection integrated in a distributed PTZ-camera network (SEMFOGO-DF)	Phase-1 analyzes panoramic image sequences to flag regions with high smoke probability. Phase-2 automatically zooms the PTZ camera and runs a smoke classifier on the zoomed view to confirm/reject the alert.	Two-phase workflow reduces false alerts from Phase-1 with a small reduction in detection rate; the system is designed for distributed, real-world operation in Brazil's Federal District.	No quantitative precision metrics provided The sequential two-phase process may increase computational overhead and response latency.
8	Y. Xu, J. Li, L. Zhang, H. Liu, and F. Zhang [30]	Deep learning (object detection; YOLOv7 enhanced)	YOLOv7 enhanced with ConvNeXtV2 backbone and CBAM attention (CNTCB-YOLOv7)	Precision 86.18%, Recall 81.85%, AP 88.36%	Detection was evaluated on static images only; no real-time video inference or latency analysis was reported.

III. RESEARCH METHODOLOGY

Deep learning for fire detection relies on large, diverse datasets. Flames vary in color (blue, red, yellow, or sometimes even green, it's different depending on the material/temperature), and smoke is gray, white, and black. Because fires occur in agriculture, infrastructure, homes, and forests, datasets must span various conditions for robust detection. Key public datasets:

BoWFires: 226 images (119 fire and 107 non-fire) at various resolutions [31].

Corsican Fire Dataset: 500 RGB fire images, 100 RGB+NIR pairs, and five RGB–NIR sequences. Each RGB image has a pixel fire mask, and NIR/sequence data mainly for multispectral analysis [32].

Smoke100k: 100,000 images with smoke/no-smoke backgrounds [33].

D-Fire: 21,000+ images in fire, smoke, fire + smoke and background classes, with YOLO-format bounding boxes [34].

VisiFire Annotations: 2,684 frames from 12 public videos, with pixel-level fire segmentation for evaluation [35].

A. Fire-Smoke 100k (FS-100k)

The FS-100k dataset offers a large-scale and diverse benchmark for combined fire–smoke detection. It contains 100,000 images sourced from both real and synthetically generated material. It covers a wide range of flame and smoke characteristics and environmental contexts.

Data Sources and Diversity: Frames were extracted at different intervals from public clips, capturing both indoors (kitchens and garages) and outdoors (wildfires and industrial sites). The dataset includes transparent smoke, flickering flames, night scenes, and off-angle viewpoints, providing a range of visual conditions that commonly occur in real situations. To represent rare or hard-to-capture scenarios, real

images were supplemented with synthetic cases created using compositing techniques and generative models.

These cases include:

- large wildfire edges
- Multi-colored flames
- heavy smoke in low light or fog

The synthetic overlays introduce additional variety in smoke density and flame appearance, helping cover situations that may be underrepresented in standard footage.

B. High-Resolution Close-UP Fire Imagery

- To capture fine-grained flame traits, the dataset included high-resolution video frames from a modern smartphone with optical and telephoto sensors. These samples show close-range fire behavior from controlled sources such as:
- Small ignition devices (matchsticks, lighters)
- Steady indoor flames under varied lighting
- Low velocity ignition sources resembling real triggers.
- This enables detailed flame representation, including:
- The Edge texture and Core brightness/ Micro flickers, Heat distortion.

C. Environmental Variation and Image Characteristics

Urban: buildings, vehicles

Rural: fields, farms

Forested: wildlands, parks

Industrial: factories, chemical plants

The weather ranges from sunny to overcast, rainy, and foggy, ensuring diverse backgrounds.

Resolution: 360p–1080p, detailed enough for small flames and faint smoke.

Flame Colors: From deep red to bright yellow, white-hot, occasional blue (gas/chemical), and multi-hued transitions.

Smoke Attributes: From transparent or light-gray plumes to thick black industrial smoke, often occluding backgrounds or overlapping flames.

Examples are shown in Fig. 1–8.

D. Data Collection and Generation

To build a diverse dataset suitable for fire and smoke detection tasks, 100,000 images were collected from multiple sources, including public footage, mobile captures, and synthetically generated samples. Data acquisition emphasized visual diversity, resolution quality, and coverage of varied real-world conditions.

Fig. 1. Indoor fire example in a living room.

Fig. 2. Indoor fire example in a kitchen.

Fig. 3. Outdoor fire example in an urban area.

Fig. 4. Outdoor fire example in a forest.

Fig. 5. Controlled flame example.

Fig. 6. Fire with dense black smoke light.

Fig. 7. Stage example under bright spotlights (non-fire light source).

Fig. 8. Close-up example of a lit matchstick.

1) Public online video footage: Part of the dataset was created by extracting frames from publicly available videos of fire and smoke incidents, including vehicle fires, wildfires, residential, and industrial events. Frames were sampled at different intervals to capture variations in flame appearance, intensity, and smoke development.

2) High-resolution fire imagery via mobile capture: To expand the dataset with both controlled and spontaneous scenes, high-resolution images were captured indoors and outdoors using mobile devices. These photos include varied ignition sources, faint smoke, and diverse lighting conditions, offering close-range details that are not commonly found in web-sourced datasets.

Examples include:

Indoor scenes (candles, cigarettes, lighters)

Outdoor scenes with subtle smoke (burning paper, small debris)

Urban and residential environments (streets, homes, rooftops) are captured at various times of day under natural and artificial lighting.

These variations help incorporate low visibility and early-stage smoke conditions that are often underrepresented in public datasets.

3) Synthetic sample generation: To cover rare or difficult-to-capture scenarios (dense smoke in confined spaces, industrial flare-ups, and flame reflections), synthetic images were generated using generative models and compositing. The selected samples were used to supplement underrepresented cases in the dataset. Each sample, regardless of origin, was manually reviewed for quality, clarity, and relevance before annotation and inclusion in the final dataset (Table II).

TABLE II. CATEGORIES AND ENVIRONMENTAL VARIATIONS

Category	Subcategories/ Variations
Fire Scenarios	- Small flames: candles, matchsticks, lighters - Medium flames: kitchen stoves, barbeques - Large fires: building fires, wildfires
Smoke Variations	-Transparent/light smoke (early ignition, cigarettes) -Dense black smoke (industrial, vehicular) - Hazy smoke under varied lighting (indoor low-light, outdoor daylight)
Background Complexity	- Urban: traffic, buildings, busy streets - Rural/Natural: forests, grasslands, mountains - Indoor: kitchens, living rooms, garages
Lighting Weather	- Lighting: daytime vs nighttime - Weather: rainy, snowy, foggy scenes

This diversity helps expose the model to a wide range of fire and smoke appearances, which can support better performance across varied scene conditions. The FS-100K dataset includes 27,846 fire-only images representing 27.85% of the dataset, 19,624 smoke-only images representing 19.62%, 27,391 images containing both fire and smoke, representing 27.39%, and 25,139 background images without visible fire or smoke, representing 25.14%. The background samples were included to improve the model's ability to distinguish real fire and smoke from visually similar non-hazardous scenes and to reduce false-positive detections.

E. Data Annotation and Labeling

Each image was manually annotated using pixel-level segmentation as illustrated in Fig. 9.

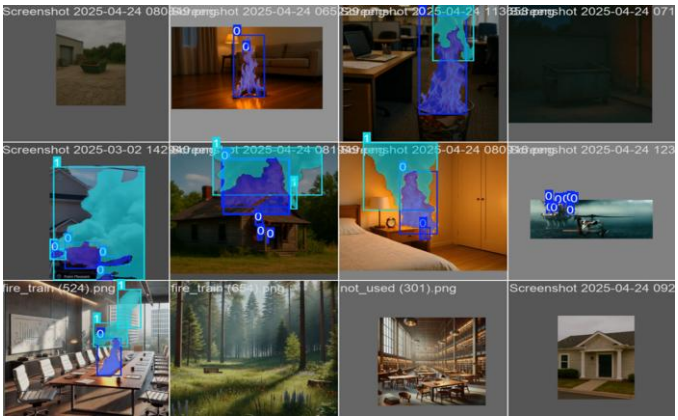


Fig. 9. Pixel-level segmentation illustrating annotated regions of fire and smoke showing fire and smoke masks under different indoor and outdoor conditions.

This fine-grained annotation supports detailed modeling of fire and smoke shapes, including transparent, thin, or partially occluded regions that are poorly represented by bounding

boxes. Segmentation masks provide more precise supervision for training both the detection and segmentation components.

Polygon manual annotation was performed, where annotators outlined fire and smoke regions to generate pixel-level masks. Manual refinement was applied for complex cases such as faint smoke, overlapping fire–smoke regions, or low-visibility scenes to maintain annotation quality across diverse conditions.

The model was trained entirely from scratch without pretrained weights, allowing the learned representations to focus specifically on fire and smoke characteristics rather than generic visual features from unrelated datasets.

To evaluate the YOLOv8-FireSmoke model, it performed an assessment on a 20K-image validation split of the FS-100k dataset. The next section presents the evaluation of the YOLOv8-FireSmoke model using both performance metrics and data analysis. In addition to standard measures such as precision–recall curves and F1-score, also analyze the distribution of bounding boxes and object locations in the dataset. The results show that most objects are concentrated near the center of the images and are generally small in size. These patterns are important because they can affect how the model learns and performs. By combining model performance with dataset analysis, this section provides a clearer and more reliable understanding of the model's behavior on the FS-100k validation set.

F. Architecture of the Model

The network consists of three main components: a backbone. The proposed model consists of a backbone for hierarchical feature extraction, a bidirectional neck for multi-scale feature fusion, and a multi-scale decoupled segmentation head for final prediction. The backbone integrates C2fFireMFA blocks to enhance feature representation using multi-scale feature aggregation (MFA), while C2fCBAM blocks are employed in deeper layers to improve attention through channel and spatial refinement. The neck adopts a top-down and bottom-up structure to effectively combine semantic and spatial information across multiple feature levels. Finally, the head operates on four scales (P2–P5) to handle objects of varying sizes, producing bounding box, classification, and segmentation outputs (Fig. 10).

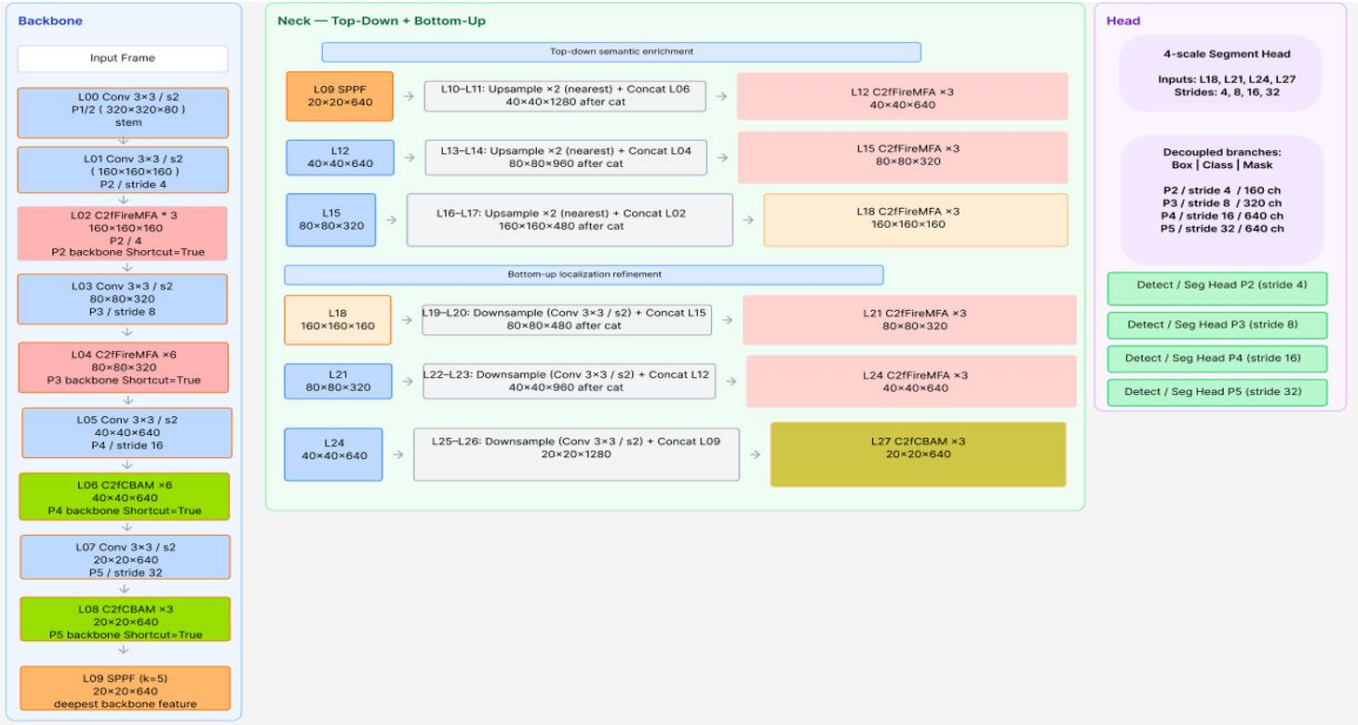


Fig. 10. Model architecture of fire-smoke.

IV. Results and Analysis

A. Accuracy Table

Table III presents a comparative summary of the key performance metrics of the proposed YOLOv8-FireSmoke model against the baseline YOLO model. The evaluation includes precision, recall, and mAP@50 measured across both fire and smoke detection tasks.

The results show an enhancement across all metrics with the Fire-Smoke model. The precision increased from 94% to 98.2% and the recall from 92% to 97.3%, indicating enhanced detection accuracy and reduced false predictions. Similarly, mAP@0.5 improved from 89% to 94%, it is reflecting better overall localization and detection performance.

These improvements were achieved with a moderate increase in model complexity, as seen in the higher number of layers (232 to 337) and parameters (71.75M to 73.53M), along with a slight rise in computational cost (GFLOPs). Overall, the proposed model achieves a balance between accuracy and efficiency, making it reliable for fire and smoke detection applications and providing a controlled evaluation under the same dataset and evaluation protocol, the original YOLOv8x

baseline and the proposed Fire-Smoke model were trained and evaluated using the same FS-100K training and validation splits (Fig. 11). Table III reports the comparison in terms of model complexity and detection performance. This controlled setting allows a direct assessment of the performance improvement achieved by the proposed architectural modifications, as shown in Table III.

TABLE III. CONTROLLED COMPARISON BETWEEN YOLOV8X BASELINE AND THE PROPOSED FIRE-SMOKE MODEL ON THE FS-100K VALIDATION SPLIT

Model	Baseline Of Yolo	Fire-Smoke Model
Layer	232	337
Parameters	71,752,774	73.53M
Precision	94%	98.2%
Recall	92%	97.3%
Map@0.5	89%	94%
GFLOPs	329	333.3

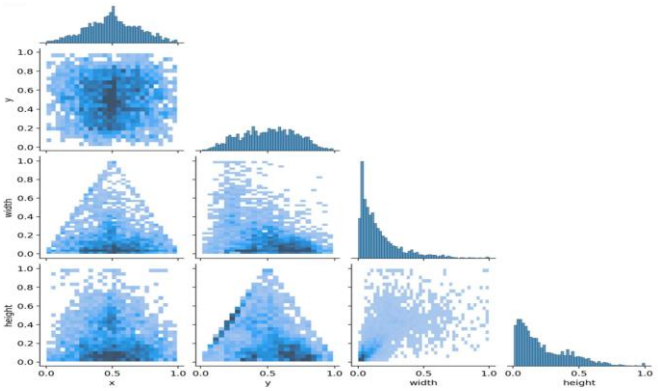


Fig. 11. Joint distributions of object center coordinates and bounding box sizes, highlighting spatial concentration and scale variability across the dataset.

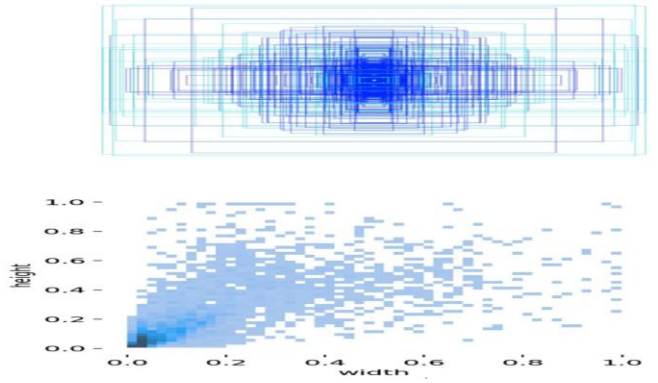


Fig. 12. Statistical analysis of annotated bounding boxes. The upper visualization depicts the spatial aggregation of bounding boxes, indicating a strong central tendency in object placement. The lower heatmap represents the joint distribution of normalized width.

As shown in Fig. 15, the precision and recall of three recent detectors are compared with the results reported in their original papers. Our YOLOv8-FireSmoke model achieves strong performance on the FS100k dataset, with a precision of 98.2% and a recall of 97.3%. Since the compared models were evaluated on different datasets, the comparison provides a general context rather than a strict benchmark. The direct comparison between the model and the YOLOv8x baseline is presented separately in Table III, where both models are evaluated using the same FS-100K validation split and evaluation protocol. Even so, the proposed YOLOv8-FireSmoke model achieves performance levels that are consistent with recent high-performing fire detection models (Fig. 14).

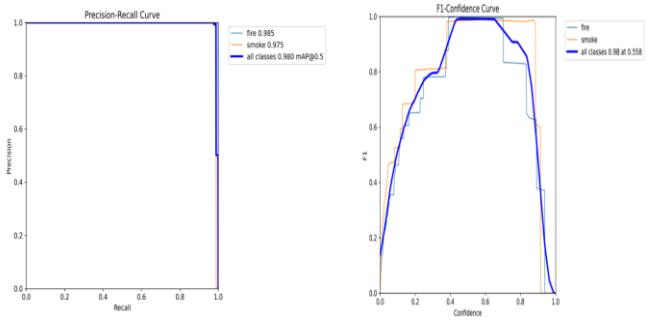


Fig. 13. Dual-metric visualization summarizing detector performance. Left: F1-score versus confidence threshold for fire, smoke, and overall classes—highlighting a peak F1 ≈98% at confidence ≈ 0.55. Right: Precision-recall curves at IoU 0.5.

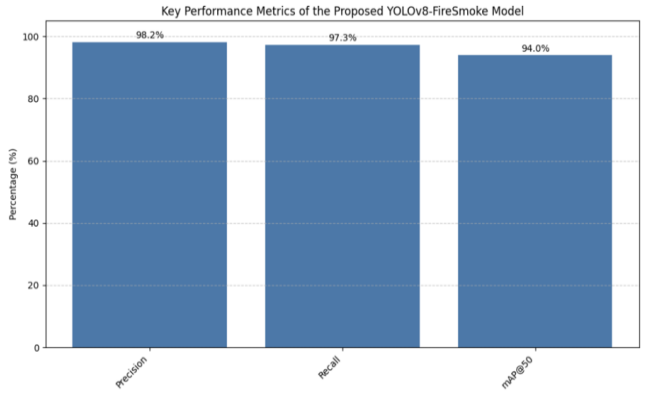


Fig. 14. Bar chart summarizing the proposed YOLOv8-FireSmoke model's precision, recall, and mAP@50 performance.

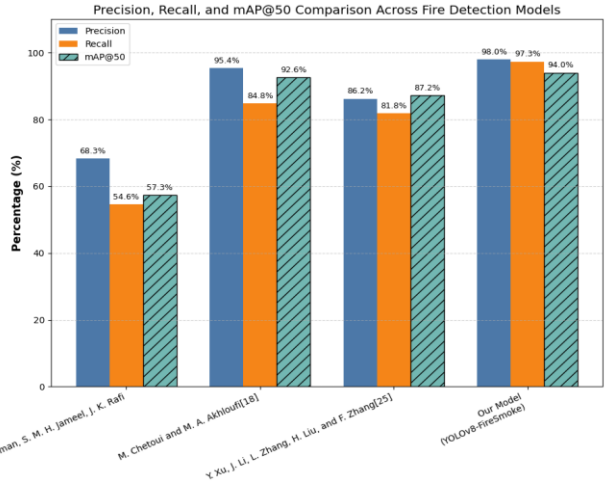


Fig. 15. Comparative evaluation of fire and smoke detection models based on precision, recall, and mAP@50.

B. Model Performance and Configuration

The results indicate that the model can localize and detect fire. The results provide both visual and numerical information about the performance of the YOLOv8-FireSmoke model. From the distribution plots (Fig. 12), object locations are more concentrated near the center of the images, while bounding box sizes vary. This suggests that the dataset contains objects at different positions and scales.

The performance curves (Fig. 13) show that the model maintains a high F1-score across different confidence levels, with a peak of about 0.98 at a confidence of around 0.55. The precision-recall curves for fire and smoke also indicate stable performance without sharp drops.

Based on Table III, the proposed model performs better than the baseline. Precision increases from 94% to 98.2%, recall from 92% to 97.3%, and mAP@0.5 from 89% to 94%. This improvement comes with a small increase in computational cost.

The training configuration was selected to support stable learning and improve the model's ability to detect fire and smoke under different visual conditions. The model was trained for 100 epochs using an input image resolution of 768 × 768 pixels, which provides sufficient spatial detail for detecting small fire and smoke regions. A batch size of 16 was used to balance training stability with memory requirements.

AdamW was selected as the optimizer because it combines adaptive gradient-based optimization with decoupled weight decay, helping to control overfitting during training. The initial learning rate was set to 1×10^{-3}, and a cosine learning-rate schedule was applied to gradually reduce the learning rate as training progressed. A 5-epoch warm-up stage was used at the beginning of training with a warm-up momentum of 0.9 and a warm-up bias learning rate of 0.01, allowing the model to start learning gradually before full optimization. The momentum parameter was set to 0.937, and the weight decay was set to 5×10^{-4}. Data augmentation included HSV color adjustment, horizontal flipping, translation, scaling, mosaic augmentation, copy-paste augmentation, and, for segmentation, overlapping masks were enabled, and the mask ratio was set to 4. The random seed was set to 42.

Overall, the results suggest that the YOLOv8-FireSmoke framework may be suitable for real-time applications such as smart building monitoring, wildfire observation systems, industrial safety, and UAV-based inspection, where timely detection is important.

C. Threat Level Algorithm

A threat level post-processing step was designed to differentiate small, controlled ignition sources from larger Dangerous fires. The algorithm evaluates factors such as flame size, surrounding smoke density, and the spatial relationship between fire and smoke. Small, localized ignition sources such as lighters, matchsticks, or brief household ignition events are treated as low threat cases, while larger fires with substantial smoke development are assigned higher threat levels.

This refinement helps reduce false alarms in scenes where minor ignition sources are visually similar to early-stage fire events.

1) Threat level post-processing: A post-processing module was introduced to differentiate small, controlled ignition sources (lighters, matchsticks, brief household ignition events) from larger fire events that exhibit substantial flame growth or smoke development. The module uses visual cues such as flame size, spatial extent, and surrounding smoke density to assign a threat category to each detected fire region. This additional step helps reduce false alarms in situations where small ignition sources may appear visually similar to early-stage fire events.

2) Step of threat level algorithm: Once a fire region is detected by the model, a secondary analysis evaluates contextual information to refine its threat category. The refinement considers attributes such as the relative size of the flame, proximity and density of visible smoke, and the spatial consistency of the fire across frames. Based on these cues, each detected fire instance is assigned to one of three levels, as illustrated in Fig. 16–19.

3) The decision is based on: The threat level estimation uses three main cues extracted from the detection output:

Fire area (number of pixels of the fire).

Smoke proximity: Distance from the fire and smoke.

Smoke Density.

4) Parameter definitions: The T-safe fire area threshold separates a small harmless flame from a large, dangerous flame.

D-crit – maximum center-to-center distance, in pixels, for treating smoke near the fire.

S-safe minimum smoke area ratio regarded as dense, indicating active combustion

D. Threat Level Rules

1) Normal fire – No threat

As shown in Fig. 16.

Fig. 16. Testing in real time for the proposed model for (normal fire).

Condition fire area < T-safe and (no smoke within D-crit or smoke ratio < S-safe).

Examples: candle flame, lighter, matchstick.

2) Dangerous fire – High threat

As shown in Fig. 17.

Fig. 17. Testing in real time for the proposed model (normal fire and dangerous fire).

Condition: fire area ≥ T-safe and (no smoke within D-crit or smoke ratio < S- safe).

Examples: stove flare-up, unattended large flame.

3) Real Threat

As shown in Fig. 18-19.

Fig. 18. Testing in real time for the proposed model (real threat)(a).

Fig. 19. Testing in real time for the proposed model (real threat)(b).

Condition: centroid distance between fire and smoke ≤ D-crit, and either fire area ≥ T-safe or smoke area ≥ T-safe. Examples: small ignition producing thick smoke, large fire with nearby smoke, spreading indoor fire.

As shown in: https://youtu.be/JXU0-0EI7Ns

The threat-level classification in the current experimental configuration was based on the mask area of the detected fire and smoke regions and on the spatial relationship between them. The safe area threshold, T-safe, was set to 1000 mask pixels. Therefore, any isolated fire or smoke region with a mask area below 1000 pixels was treated as a small, low-risk detection, while any isolated fire or smoke region with a mask area equal to or greater than 1000 pixels was treated as a dangerous, high-risk detection. The critical distance threshold, D-crit, was set to 90 pixels. This value was used as the maximum centroid-to-centroid distance for considering a fire region and a smoke region spatially associated. When fire and smoke were detected within this distance, they were treated as a related fire-smoke pair. If both regions in the pair had mask areas below T-safe, the case was classified as a normal low-risk fire-smoke detection. However, if the fire and smoke regions were spatially associated and one of them had a mask area equal to or greater than T-safe, the case was classified as a real threat. Before threat classification, same-class detections were merged. Fire regions were merged with fire regions, and smoke regions were merged with smoke regions, when their IoU was 0.5 or higher or when the centroid distance between them was 30 pixels or less. The final threat level was then assigned after calculating the merged mask areas and centroid distances. If multiple detections were present in the same frame, the highest detected threat level was assigned to the frame. The FS-100K dataset, containing over 100,000 images, is available from the corresponding author upon reasonable request.

E. Threat Level Algorithm Flowchart (Fig. 20)

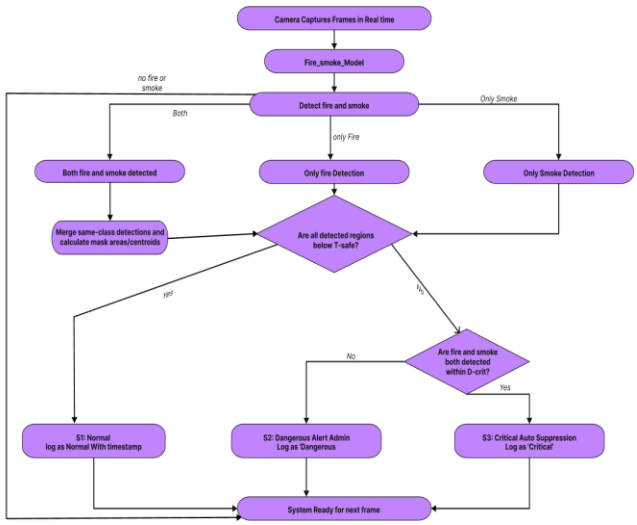


Fig. 20. An overview of the proposed intelligent fire detection threat level algorithm.

V. CONCLUSION

This work presented YOLOv8-FireSmoke, a real-time deep learning model for fire detection. The approach integrates three main components:

An optimized architecture consisting of a backbone, neck, and multi-scale detection head. The backbone is responsible for hierarchical feature extraction, while the neck employs a top-down and bottom-up feature aggregation strategy to enhance semantic and spatial information across multiple scales. The detection head further processes these features to enable accurate localization.

Three-class post-processing logic that separates normal ignition sources, dangerous fires, and fires–smoke threats. This

step helps reduce false alarms while ensuring that hazardous events are escalated quickly.

FS-100K dataset, consisting of 100,000 diverse images from real and synthetic sources and covering indoor, outdoor, low light, daylight, and challenging edge case scenarios. On a 20,000-image validation split, the proposed model achieved 94% mAP@50 and precision above 98%.

Real-time inference speed, processing video at approximately 57 FPS on a single RTX 4070 Super GPU, including post-processing. This performance exceeds the requirements of real-time surveillance and early warning systems.

Overall, the results show that combining rich training data with context-sensitive post-processing yields an accurate and practical framework suitable for deployment in urban, industrial, and wildland environments. Vision-based systems also offer advantages over fixed heat or smoke sensors, including earlier visual response and fewer nuisance alarms in complex spaces. As fire risk increases in many regions, deep learning–based detection may support the transition from reactive response to more proactive monitoring. Future work may incorporate thermal or multispectral inputs to further improve robustness in extremely low visibility conditions.

VI. Future Work

Multispectral and Thermal Fusion.

Integrate infrared or thermal cameras with RGB streams to improve detection at low visibility in nighttime or very hard smoke conditions in darkness.

3D Localization and Multi Camera Geometry.

Employ 3D bounding boxes or stereo/multi-camera rigs to refine distance and volumetric growth estimates, enabling better threat assessment and resource allocation.

Predictive spread Modelling decouples the visual detector with atmospheric and fuel load data to forecast the likely spread direction or growth rate, turning early warning into actionable predictions.

References

[1] O. Haas, I. C. Prentice, and S. P. Harrison, “Wildfires on a changing planet,” Nat. Commun., vol. 17, p. 1599, 2026, doi: 10.1038/s41467-025-68176-4.

[2] C. X. Cunningham et al., “Wildfires will intensify in the wildland–urban interface under near-term warming,” Commun. Earth Environ., vol. 6, p. 542, 2025, doi: 10.1038/s43247-025-02475-y.

[3] D. I. Kelley et al., “State of Wildfires 2024–2025,” Earth Syst. Sci. Data, vol. 17, pp. 5377–5488, 2025, doi: 10.5194/essd-17-5377-2025.

[4] S.-Y. Kim and A. Muminov, “Forest fire smoke detection based on deep learning approaches and unmanned aerial vehicle images,” Sensors, vol. 23, no. 12, p. 5702, 2023, doi: 10.3390/s23125702.

[5] J. Sandoval Díaz et al., “Community resilience to wildfires: A systematic review of impacts, coping strategies, indicators, and governance challenges,” Prog. Disaster Sci., vol. 27, p. 100447, 2025, doi: 10.1016/j.pdisas.2025.100447.

[6] Z. A. Mani, A. Khorram-Manesh, and K. Goniewicz, “Global health impacts of wildfire disasters from 2000 to 2023: A comprehensive analysis of mortality and injuries,” Disaster Med. Public Health Prep., vol. 18, p. e230, 2024, doi: 10.1017/dmp.2024.150.

[7] S. P. Naik, V. Lohbande, S. Hambir, R. Korade, and R. Hatkar, “Fire detection with image processing,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 11, no. 5, pp. 2123–2128, 2023, doi: 10.22214/ijraset.2023.52073.

[8] C. Jin, T. Wang, N. Alhusaini, S. Zhao, H. Liu, K. Xu, and J. Zhang, “Video fire detection methods based on deep learning: Datasets, methods, and future directions,” Fire, vol. 6, no. 8, p. 315, 2023, doi: 10.3390/fire6080315.

[9] J. Carnicer, A. Alegría, C. Giannakopoulos, F. Di Giuseppe, A. Karali, N. Koutsias, P. Lionello, M. Parrington, and C. Vitolo, “Global warming is shifting the relationships between fire weather and realized fire-induced CO_2 emissions in Europe,” Sci. Rep., vol. 12, p. 10365, 2022, doi: 10.1038/s41598-022-14480-8.

[10] S. McElroy, L. Schwarz, H. Green, I. Corcos, K. Guirguis, A. Gershunov, and T. Benmarhnia, “Defining heat waves and extreme heat events using sub-regional meteorological data to maximize benefits of early warning systems to population health,” Sci. Total Environ., vol. 721, p. 137678, 2020, doi: 10.1016/j.scitotenv.2020.137678.

[11] A. Salaria and A. Singh, “A guide for selection of wireless communication technology for effective and robust early forest fire detection system,” Bull. Electr. Eng. Inform., vol. 14, no. 2, pp. 1026–1035, 2025, doi: 10.11591/eei.v14i2.8613.

[12] M. Bugarić, D. Krstinić, L. Šerić, and D. Stipanicev, “Current trends in wildfire detection, monitoring and surveillance,” Fire, vol. 8, no. 9, p. 356, 2025, doi: 10.3390/fire8090356.

[13] S. Khan and A. Khan, “FFireNet: Deep learning-based forest fire classification and detection in smart cities,” Symmetry, vol. 14, no. 10, p. 2155, 2022, doi: 10.3390/sym14102155.

[14] B. M. Diaconu, “Recent advances and emerging directions in fire detection systems based on machine learning algorithms,” Fire, vol. 6, no. 11, p. 441, 2023, doi: 10.3390/fire6110441.

[15] J. Pesonen et al.,Detecting wildfires on UAVs with real-time segmentation trained by larger teacher models, in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2025.

[16] P. Barmpoutis, P. Papaioannou, K. Dimitropoulos, and N. Grammalidis, “A review on early forest fire detection systems using optical remote sensing,” Sensors, vol. 20, no. 22, p. 6442, 2020, doi: 10.3390/s20226442.

[17] N. D. Ismail, R. Ramli, and M. N. A. Rahman, “A systematic literature review of vision-based fire detection, prediction and forecasting,” Jurnal Kejuruteraan, vol. 37, no. 1, pp. 191–218, 2025, doi: 10.17576/jkukm-2025-37(1)-14.

[18] S. A. Elhanashi, S. Essahraui, P. Dini, and S. Saponara, “Early fire and smoke detection using deep learning: A comprehensive review of models, datasets, and challenges,” Appl. Sci., vol. 15, no. 18, p. 10255, 2025, doi: 10.3390/app151810255.

[19] L. Deng, S. Wu, S. Zou, and Q. Liu, “Large-space fire detection technology: A review of conventional detector limitations and image-based target detection techniques,” Fire, vol. 8, no. 9, p. 358, 2025, doi: 10.3390/fire8090358.

[20] U. U. Deshpande et al., “Real-time fire and smoke detection system for diverse indoor and outdoor industrial environmental conditions using a vision-based transfer learning approach,” Front. Comput. Sci., vol. 7, p. 1636758, 2025, doi: 10.3389/fcomp.2025.1636758.

[21] L. He, Y. Zhou, L. Liu, Y. Zhang, and J. Ma, “Research and application of deep learning object detection methods for forest fire smoke recognition,” Sci. Rep., vol. 15, p. 16328, 2025, doi: 10.1038/s41598-025-98086-w.

[22] A. Abozeid and R. Alanazi, “An intelligent approach for early smoke/fire detection using vision sensors in smart cities,” Sci. Rep., vol. 16, p. 11387, 2026, doi: 10.1038/s41598-026-42762-y.

[23] S. Danish et al., “Vision-based fire management system using autonomous unmanned aerial vehicles: A comprehensive survey,” Artif. Intell. Rev., vol. 59, p. 16, 2026, doi: 10.1007/s10462-025-11415-3.

[24] M. Chetoui and M. A. Akhloufi, “Fire and smoke detection using fine-tuned YOLOv8 and YOLOv7 deep models,” Fire, vol. 7, no. 4, p. 135, 2024, doi: 10.3390/fire7040135.

[25] M. Prakash, S. Neelakandan, M. Tamilselvi, S. Velmurugan, S. B. Priya, and E. O. Martinson, “Deep learning-based wildfire image detection and

classification systems for controlling biomass," Int. J. Intell. Syst., vol. 2023, Art. no. 7939516, pp. 1–18, 2023, doi: 10.1155/2023/7939516.

[26] R. P. Vignesh, P. Darshan Kumar, S. Rajasekar, and K. Prakash, "Fire and smoke detection using the YOLO algorithm," Int. J. Mod. Eng. Res., vol. 14, no. 3, pp. 26–30, 2024.

[27] S. Rahman, S. M. H. Jamee, J. K. Rafi, J. S. Juthi, S. A. Aziz, and J. Uddin, "Real-time smoke and fire detection using You Only Look Once v8-based advanced computer vision and deep learning," Int. J. Adv. Appl. Sci., vol. 13, no. 4, pp. 987–999, 2024, doi: 10.11591/ijaas.v13.i4.pp987-999.

[28] M. Mukhiddinov, A. B. Abdusalomov, and J. Cho, "A wildfire smoke detection system using unmanned aerial vehicle images based on the optimized YOLOv5," Sensors, vol. 22, no. 23, p. 9384, 2022, doi: 10.3390/s22239384.

[29] N. Borges et al., "A fire management intelligent system for the Brazilian Cerrado biome based on a deep learning two-phase detection method," J. Reliab. Intell. Environ., vol. 11, no. 1, p. 5, 2025, doi: 10.1007/s40860-025-00244-4.

[30] Y. Xu, J. Li, L. Zhang, H. Liu, and F. Zhang, "CNTCB YOLOv7: An effective forest fire detection model based on ConvNeXtV2 and CBAM," Fire, vol. 7, no. 2, p. 54, 2024, doi: 10.3390/fire7020054.

[31] D. Y. T. Chino, L. P. Avalhais, J. F. Rodrigues Jr., and A. J. M. Traina, "BoWFire dataset: Fire and non-fire still images for pixel-based detection," Univ. of São Paulo (GBDI-ICMC). [Online]. Available: https://bitbucket.org/gbdi/bowfire-dataset/downloads/

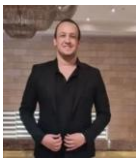
[32] Environmental Sciences UMR CNRS 6134 SPE Lab., "Corsican Fire Database," Univ. of Corsica Pasquale Paoli. [Online]. Available: https://cfdb.univ-corse.fr/

[33] H.-Y. Cheng, J.-L. Yin, B.-H. Chen, and Z.-M. Yu, "Smoke100k: A database for smoke detection." [Online]. Available: https://github.com/BigMM/Smoke100k

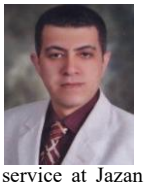
[34] D-Fire, "An image dataset for fire and smoke detection," GitHub repository. [Online]. Available: https://github.com/gaiasd/DFireDataset

[35] W. Wahyono, A. Harjoko, A. Dharmawan, F. D. Adhinata, G. Kosala, and K. H. Jo, "Region-based annotation data of fire images for intelligent surveillance systems," Zenodo, 2022, doi: 10.5281/zenodo.5893854.

AUTHORS' PROFILE



Mohamed Youssef, He earned his B.Sc. in Computer Science and AI from MTI University and is currently pursuing an M.Sc. in Computer Science, specializing in Data Science. He also holds a Master of Business Administr action (MBA) from New Mexico and a professional cybersecurity (CC) certificate. His background spans data science, AI, and cybersecurity, with research interests in deep learning, computer vision, and AI applications.



MOHAMED I. MARIE is a Professor of Information Systems at the Faculty of Computers and Artificial Intelligence, Capital University. He received his B.Sc. degree from Cairo University in 1994, followed by his M.Sc. and Ph.D. degrees from Helwan University in 1999 and 2004, respectively. He has held a range of academic and administrative positions throughout his career, including his service at Jazan University from 2012 to 2019. His scholarly contributions include more than 30 published research papers, in addition to reviewing over 70 manuscripts for academic venues. He also maintains research profiles and academic records on several platforms, including Publons, Web of Science, ORCID, Academia, and Google Scholar.

Sarah Naiem, she is a faculty member at the Faculty of Computer Science and AI, Capital Unive rsity, where she lectures and conducts research in computer science and information systems. Her interests include cloud computing, data science, text and opinion mining, software engineering, and information systems.

A Novel Hybrid Deep Learning Validation Model for Real-Time and Synthetic Image Inputs in Capsicum Plant Disease Diagnosis

Prashant Vikhe[1], Baisa Gunjal[2]
MET's Institute of Engineering, India[1]
Amrutvahini College of Engineering, India[2]

***Abstract*—The diagnosis of plant diseases in Capsicum species remains a critical challenge in precision agriculture due to variability in environmental conditions and limited availability of high-quality datasets. The traditional convolutional neural network (CNN) has been demonstrated to have satisfactory performance, but it cannot capture robust performance in both real-time and synthetic images. This study introduces a novel hybrid deep learning validation model based on Convolutional neural networks (CNN) combined with Capsule networks (CapsNet) and Vision Transformer (ViT) backbone. The framework is intended to be able to validate multi-source image inputs and improve the reliability of the classification in natural and synthetic image environments. In contrast to previous plant disease detection models, which can only be trained on real-time plant images, the proposed CNN–ViT–CapsNet framework features a dual-domain validation process that is able to classify both real-time and the GAN-generated synthetic Capsicum leaf images. CNN-based local feature extraction, ViT-based global contextual learning, and CapsNet-based spatial validation ensure robustness against illumination, orientation, and background conditions.**

***Keywords*—Capsicum plant disease; hybrid deep learning; real-time validation; synthetic images; image processing; CNN-CapsNet-ViT**

I. Introduction

Crop diseases are a major threat to agriculture and sustainable food production worldwide, with early diagnosis and management being essential for the sustainability of food production. The capsicum species, such as chili pepper and bell pepper, are economically important horticultural crops grown all over the world for their food, spice, and pharmaceutical value. They are, however, very susceptible to fungal, bacterial, and viral diseases, including Cercospora leaf spot, anthracnose, bacterial wilt, viral mosaic infections, etc. Besides affecting crop yield, these diseases diminish fruit quality and result in economic losses for producers and disruption in the food value chain. The traditional diagnostic approaches of manual visual inspection by plant health experts, or laboratory assays, are labor-intensive, time-consuming, and may not be suitable for in-field use. This has made automated, intelligent, and scalable disease detection systems vital to ensure plant health and sustainable agriculture.

Significant progress has been made in recent years in the area of detection of plant diseases via artificial intelligence (AI), computer vision, and deep learning (DL). Convolutional Neural Networks (CNNs) have shown promising performance for feature extraction, while new architectures such as Vision Transformers (ViTs) and Capsule Networks (CapsNets) have proved to be more effective at understanding the global context and at providing orientation-independent validation for images. However, the majority of existing models are trained on small-scale real-time image datasets that are captured in laboratory conditions. This is a significant limitation because deep learning models are known to generalize poorly when presented with synthetic images (created through augmentation or generative adversarial networks) or real-world images that exhibit variations in lighting, occlusion, and background noise.

Hybrid deep learning approaches have emerged as a promising solution to address these limitations by combining the strengths of multiple architectures. For instance, CNNs are able to learn local texture-based features, Vision Transformers are able to learn long-range dependencies, and Capsule Networks are better at preserving the spatial hierarchy and verifying the orientation of local features than traditional pooling operations. However, despite numerous studies on hybrid architectures for plant disease detection, there remains a significant research gap in validating models across both real-time and synthetic image domains, particularly in the context of Capsicum plant pathology.

To overcome this, this study introduces a novel hybrid CNN–ViT–CapsNet model that aims to improve the performance of the classification of disease by simultaneously validating the features in the image from various input sources. The model extracts low-level features from raw images using a CNN backbone, embeds global context with the Vision Transformer (VT) encoder, and validates the spatial relationship with the Capsule Network (CN). These representations, combined with the proposed model, are more robust and capable of overcoming overfitting and generalization on real-time and synthetic images. The proposed framework can be incorporated into mobile apps, drone-based crop monitoring systems, smart greenhouse precision agriculture, and AIoT-driven agricultural platforms to enable early detection of diseases and minimize crop losses.

To sum up the highlights of this study, they may be put as follows:

- A novel hybrid CNN–ViT–CapsNet framework is designed and developed to address the robust Capsicum disease detection on real-time and synthetic images.
- Integration of a new validation approach for disease classification confidence, which makes use of the comprehensive, complementary local, global, and spatial feature hierarchies.
- Empirical validation is carried out by testing the methodology on real-world data from Capsicum leaves, synthetically generated images from Capsicum using GAN, and using state-of-the-art baselines to obtain better accuracy and robustness.
- The proposed framework scales well and can be extended to other crop species for disease diagnosis.

The remainder of this study is organized as follows: Section II reviews the existing literature. Section III presents the proposed CNN–ViT–CapsNet architecture. Section IV describes the experimental setup and discusses the obtained results. Finally, Section V concludes the study and outlines future research directions.

II. Related Work

Over the recent years, deep learning detection of plant diseases has been widely studied, and most models rely on CNN-based architectures. To illustrate, Kursun and Koklu [1] brought a hybrid ViT-AlexNet to the classification of chili leaf disease and showed that the transformer-based global attention model plus CNN local features extraction can be used to improve the recognition ability. Although Kursun and Koklu [1] achieved promising results using ViT-AlexNet, their framework was evaluated only on real-time images, limiting its applicability to synthetic data environments. In the same manner, Mathew et al. [2] used a depth-wise separable VGG19 together with Capsule Networks (CapsNet) to classify bell pepper and grape leaf diseases and indicated better performance as a result of the capability of the capsule network to maintain spatial hierarchies. De Brito et al. [3] also developed in this field by creating deep learning models to measure the severity of Cercospora leaf spot in chili peppers in real time, which can be used to monitor the progression of the disease in natural environments.

Other papers have suggested combining hybrid methods that will use CNNs with other layers next, and verify the context. Yağ and Altan [4] designed an effective hybrid AI model of agricultural environments in real-time and integrated classical machine learning with CNN-based feature extraction, but did not validate the approach on Capsicum specifically. Ilyas et al. [5] suggested DIANA, a deep learning-based paprika disease and pest phenotyping system with severity analysis, where multimodal detection is essential, but the large-scale performance becomes limited. Lee et al. [6] used deep learning along with image captioning to diagnose crop disease, in this case, red pepper data but the efficiency was still a problem in real-time implementations. The use of deep learning has greatly improved the accuracy of plant disease diagnosis systems.

Ahuja et al. [7] introduced a Vision Transformer–Capsule Network (ViCapsNet) architecture for assessing the quality of oranges and detecting diseases. Their study demonstrated the effectiveness of combining transformer-based feature extraction with Capsule Networks for capturing spatial relationships in agricultural images. However, the framework was developed for fruit inspection and was not evaluated on leaf disease datasets involving multiple disease categories. Kumar and Gunasekar [9] introduced a deep learning-based approach for fig leaf disease prediction and severity estimation. Their work highlighted the potential of Convolutional Neural Networks for disease classification and severity analysis. Nevertheless, the model primarily focused on disease identification and did not explore hybrid architectures capable of leveraging both local and global feature representations.

Other scientists concentrated on the combination of synthetic information to enhance generalization. Lakshmi et al. [10] used generative adversarial networks (GANs) to generate synthetic images to train deep learning models in agriculture, and demonstrated that they have the potential to improve the diversity of the dataset, but with limited testing of hybrid models. Pintus et al. [8] explored the AIoT applications in real-time agricultural classification, in which synthetic and real pictures were analyzed by edge computing platforms, focusing on the use of IoT and not on pathology diagnosis. On the contrary, Nagpal et al. [15] tested a hybrid deep learning system trained on real and synthetic data of wheat and barley, which overcomes the problem of domain adaptability, but their system was not applied to Capsicum.

Recent advancements have incorporated attention mechanisms and ensemble hybrids for Capsicum and related crops. Abubeker et al. [11] designed a depth-wise separable VGG19-CapsNet for bell pepper and grape leaves, achieving high accuracy but encountering overfitting on limited datasets. Dolatabadian et al. [12] explored a hybrid CNN–VGG19 model for detecting viral infections in Capsicum and grapes, confirming the utility of multi-branch models but lacking synthetic image evaluation. Mathew et al. [13] developed a GPU-assisted DSC-TransNet for handheld devices, which provided real-time classification of pepper and tomato diseases but required expensive hardware. Begum and Syed [14] proposed a hybrid system for pepper and corn classification with hybrid optimization and achieved excellent classification accuracy with a high amount of data.

There are several works in the literature that point out the shortcomings of CNN-only or hybrid architectures. Although the model used ResNet with VGG16, the accuracy was high; it was not integrated into the transformer, and it was used for the sweet pepper disease [16]. Rezk et al. [17] combined CNN and Conditional Random Fields (CRFs) for IoT applications in the agricultural field, but their system was not very robust in the presence of varying image scenarios. Chug et al. [18] proposed a hybrid system for tomato disease classification that was tested on real-time datasets, with no tests conducted using the synthetic datasets. Gupta et al. [19] studied transfer learning approaches with modified CNN architectures for tomato leaf disease classification. The study showed that using a pretrained model could be beneficial for classification and that transfer learning has potential to be used in agricultural applications. While the

method resulted in encouraging performance, it was primarily based on convolutional feature extraction with limited use of transformer-based contextual learning or capsule-based spatial validation mechanisms. Mathew et al. [20] introduced a CapsNet-based hybrid approach for the classification of Bell pepper leaf diseases. They found Capsule Networks to work well in maintaining hierarchical and spatial relationships among disease features. The attention models used were not transformer-based, which are well-suited to modeling long-range dependencies and global contextual information within a complex leaf image.

From this review, it is evident that hybrid deep learning models outperform standalone architectures in most agricultural applications. However, most existing works either (1) focus on specific real-time datasets without evaluating performance on synthetic data, or (2) lack an integrated validation mechanism that combines CNN feature extraction, ViT-based global attention, and CapsNet spatial verification. The current studies are mainly on spatial validation or local feature extraction separately. The validation of a comprehensive and coupled heterogeneous image domain is scarcely investigated. Setting a critical research gap, there is no common framework developed, and only a few have been designed specifically to check both real-time and synthetic image inputs for Capsicum disease detection to ensure reliable classification in different farming conditions. In this work, the CNN-ViT-CapsNet hybrid model will directly tackle this issue by integrating complementary architectures into a unified robust validation pipeline.

III. Proposed Methodology

We introduce a hybrid deep learning model combining Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Capsule Networks (CapsNets) to tackle the problems of reliable Capsicum plant disease detection in real-time images and under simulated image conditions. They are unique in their strengths and help to mitigate the weaknesses of individual stand-alone or hybrid models. Fig. 1 presents the Capsicum disease diagnosis system.

A. CNN Layers: Local Feature Extraction

The CNN-Layered is a model that extracts localized features. It is a model for local feature extraction.

CNNs are known to be effective feature extractors in image-based tasks. Here CNN is used as a back bone to extract low and mid-level spatial features of a diseased and normal Capsicum leaves, such as color distribution, texture, local edge. This is done by convolutional layers followed by pooling operations, which compress the dimensions and noise, leaving a small, yet highly discriminative representation of local disease symptoms, including leaf spots, discoloration or necrotic lesions. This step allows the model to acquire very detailed characteristics essential for identifying the fine-grained disease patterns.

The first part of the proposed framework is a Convolutional Neural Network (CNN) to learn discriminative local features from Capsicum leaf images. The CNN module consists of three layers of convolution, arranged in a hierarchy. The first convolutional layer has 32 3×3 kernels and Rectified Linear Unit (ReLU) activation function which are used to identify low level visual features such as edges and textures. The second convolutional layer is designed to enhance the ability of the model to extract features, using 64 3×3 kernels, and ReLU as the activation function, allowing the model to learn more complex patterns related to the diseases. The max-pooling operation is used to decrease spatial size, keeping the most informative features. Then, a third convolutional layer with 128 kernels further processes the extracted representations before feeding them into the transformer part. This progressive feature extraction strategy enhances the model's ability to identify subtle visual symptoms associated with different Capsicum leaf diseases.

B. ViT Encoder: Global Context Modeling

The localized receptive field of CNNs is useful for capturing the local features but cannot model long-range dependencies. To cope with this, we use a Vision Transformer (ViT) encoder. The ViT module segments outputs of CNN into patches, adds positional encoding, and applies multi-head self-attention to capture the global spatial context of the entire leaf image and to model the correlations among all patches. This enables the model to learn about characteristics of disease progression, background variations in context, consistency of structures for synthetic vs. real time inputs. The ViT has been designed to be robust to disease classification in diverse field conditions by embedding global information.

The CNN module outputs feature maps which are broken into 16×16 image patches without overlap. These patches are then converted into a representation known as "vector embeddings", which are 768-dimensional and include positional information to maintain spatial relationships. The transformer encoder is formed by twelve stacked layers of multi-head self-attention with 12 attention heads. The ViT module learns contextual relationships between the image regions by self-attention operations, which enables the effective modeling of the patterns of a disease spread across the entire leaf surface. Contextual learning with transformer improves the robustness of the system to illumination, orientation and background changes.

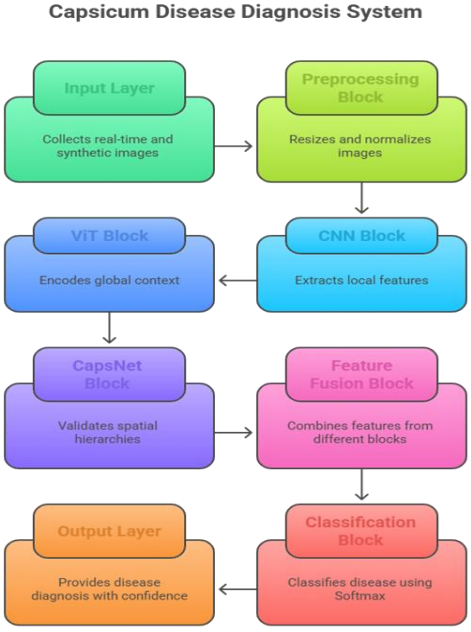


Fig. 1. Capsicum disease diagnosis system.

C. Capsule Network: Spatial Validation and Orientation Invariance

Conventional CNN pooling layers tend to ignore the spatial information, which results in the loss of important details when the disease symptoms occur with different orientations, scales, or lighting. For this reason, we overcome it by adding a Capsule Network (CapsNet). Capsules are vector-based neurons that can hold hierarchical part-to-whole relationships in image structures. This study presents CapsNet, a method to validate the spatial hierarchies of extracted features, which is designed to reliably detect disease symptoms (such as lesions, blight patches) despite their rotation, scale, and distortions in the image. The dynamic routing mechanism in CapsNet further bolsters the strength of the classification decisions, resulting in the model being very generalizable in diverse image inputs.

The capsule layer consists of 32 basic capsules that encode the basic features found in the previous layers. The capsules are linked to a higher-level capsule layer with 16-digit capsules, in which each capsule is a more abstract representation of a feature of a particular disease. Dynamic routing is done for 3 iterations to determine the strength of the connections between the lower-level and higher-level capsules. The routing mechanism allows the model to preserve spatial and structural details, which can be useful in traditional deep learning models. The proposed framework integrates the capsule-based feature validation with CNN and transformer representations, leading to a robust and accurate disease classification for both real-time and synthetic Capsicum leaf images.

D. Fusion and Classification

The outputs of CNN, ViT encoder and CapsNet are concatenated in a feature-level fusion and passed into fully connected layers. The class label and a probability score of the disease are predicted by a Softmax classifier. This fusion method enables the network to benefit from the local sensitivity of CNNs, the global context-awareness of ViTs and the orientation-invariance of CapsNets all at once.

IV. Experimental Setup and Results

For this study, two types of image datasets were used.

A. Dataset Description

1) Real-time dataset: High-resolution mobile cameras were used to collect data from Capsicum cultivation fields. Contains 5,000 images of leaves showing common diseases like Cercospora Leaf Spot, Anthracnose, Bacterial Wilt, Mosaic Virus, and Healthy. Photographs were taken with different lighting, background, and occlusion conditions.

2) Synthetic dataset: Created with Generative Adversarial Networks (GANs) and high-quality augmentation methods. Contains 2,000 synthetic images for the same classes of disease. Represents and generalizes well. The whole data set (7,000 images) was divided into three parts, with 70% for training, 15% for validation, and 15% for testing.

B. Experimental Setup

- Hardware: NVIDIA RTX 4090 GPU, Intel i9 processor, 64 GB RAM.
- Software: TensorFlow 2.13, Python 3.10, CUDA 12.
- Input size: 224×224×3
- Batch size: 32.
- Optimizer: Adam (learning rate = 0.0001).
- Loss function: Categorical Cross-Entropy.
- Epochs: 100 with early stopping (patience = 10).
- Metrics: Accuracy, Precision, Recall, F1-score, AUC (Area Under ROC Curve).

C. Quantitative Results

In the results, it is shown that the proposed hybrid is better than the traditional and hybrid baselines on real-time and synthetic datasets (see Table I).

TABLE I. Performance Comparison of Proposed Model vs. Baselines.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
CNN	89.5	88.7	88.1	88.4	90.2
ResNet50	91.2	90.5	90.0	90.2	92.1
VGG19	90.8	90.1	89.4	89.7	91.5
ViT	93.1	92.4	91.8	92.1	94.0
CapsNet	92.7	92.1	91.6	91.8	93.3
CNN–CapsNet	94.2	93.6	93.1	93.3	94.8
CNN–ViT	95.3	94.7	94.0	94.3	95.6
Proposed CNN–ViT–CapsNet	**96.8**	**96.2**	**95.8**	**96.0**	**97.5**

D. Analysis of Real-Time versus Synthetic Inputs

- The proposed model was robust even with field variability and was efficient in real-time images with 96.5% accuracy.
- It was found to be very adaptable to GAN-augmented inputs on synthetic images, with an accuracy of 97.1%.
- This means that the hybrid approach proved to be a good approach to connect real and synthetic data sets, better than models trained using only one type of data.

E. ROC Curve and Confusion Matrices

- The proposed hybrid model outperforms CNN (AUC: 90.2%) and ResNet50 (AUC: 92.1%) with an AUC of 97.5% as displayed in the ROC curve.
- The confusion matrix pointed out less misclassification between Anthracnose vs. Cercospora in CNN and VGG19.

F. Discussion

The proposed CNN–ViT–CapsNet model achieves better performance than the other models due to three factors:

1) Local–Global–Spatial synergy: CNN captured fine-grained disease spots, ViT modeled global relationships across leaf structures, and CapsNet validated orientation-invariant features.

2) Synthetic data robustness: The proposed architecture was generally able to generate good results in both real and synthetic data sets.

3) Feature fusion strategy: Joint feature representation led to complementary strengths being utilized, which led to greater classification confidence.

The model has high accuracy, but its computation is more expensive than the CNN-only model. In order to train the model efficiently, the use of a GPU is recommended. But, after training, inference speed is still appropriate for real-time agricultural applications.

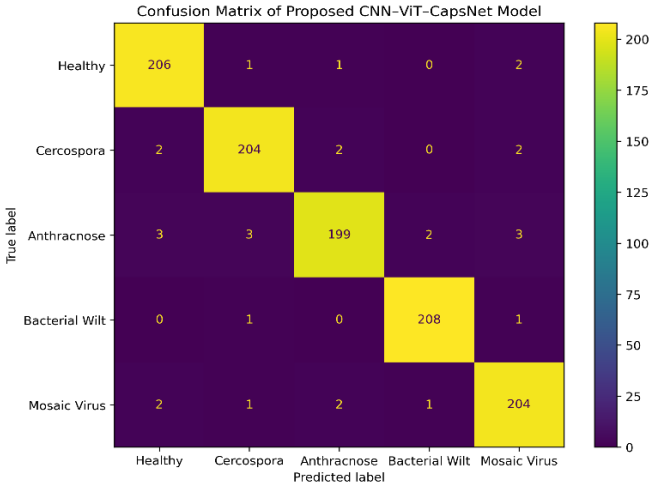


Fig. 2. Confusion matrix

Fig. 2 shows the confusion matrix of the proposed CNN–ViT–CapsNet model, giving a deep insight into the classification performance for the five classes in the target classes. The high values in the diagonal of the matrix indicate that the recognition accuracy is good for Healthy, Cercospora, Anthracnose, Bacterial Wilt, and Mosaic Virus. Misclassifications are relatively few and are spread among samples, e.g., Cercospora is sometimes misclassified as Healthy, Mosaic Virus is sometimes misclassified as Cercospora, Anthracnose is sometimes misclassified as Healthy, and Mosaic Virus is sometimes misclassified as Anthracnose. Interestingly, Bacterial Wilt has the highest performance among all the diseases, and there are very few errors, which shows the model's capability of distinguishing it well from the other diseases. Overall, the confusion matrix confirms the fact that the proposed model provides a good classification performance and has a very low mislabeling rate, thus supporting the application of the proposed model in the activity of plant disease inspection.

The proposed CNN ViT CapsNet model performance is presented in the form of Receiver Operating Characteristic (ROC) curves for five classes of Capsicum diseases, as shown in Fig. 3. The tradeoff is represented by the ROC curves between the True Positive Rate (TPR) and False Positive Rate (FPR) at varying threshold values of each type of disease. The corresponding AUC values were 0.98 for Healthy, 0.97 for Cercospora, 0.96 for Anthracnose, 0.99 for Bacterial Wilt, and 0.97 for Mosaic Virus. The proposed hybrid architecture shows high reliability to differentiate disease classes with an average AUC of around 0.975. The ROC curves are well above the random classification level, which underscores the ability of the proposed framework to effectively diagnose diseases using a combination of CNN, Vision Transformer, and Capsule Network.

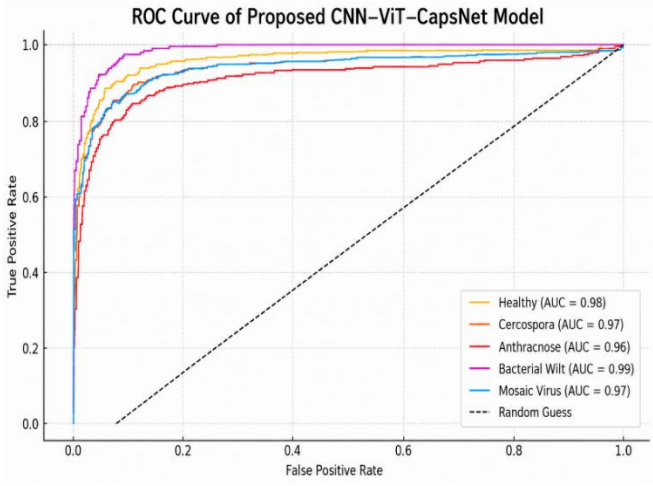


Fig. 3. ROC curve of the proposed CNN-VIT-CapsNet model.

TABLE II. CLASSIFICATION REPORT (MODEL PERFORMANCE METRICS)

Class	Precision	Recall	F1-Score
Healthy	0.97	0.96	0.96
Cercospora	0.96	0.95	0.95
Anthracnose	0.95	0.95	0.95
Bacterial Wilt	0.98	0.97	0.97
Mosaic Virus	0.95	0.96	0.95

The performance of the proposed CNN–ViT–CapsNet model is presented in Table II in terms of its class-wise performance. The results demonstrate a very high level of classification accuracy for all the categories of the disease, with precision, recall, and F1 Score above 0.95 in most of the classes. The precise classification was highest for the Bacterial Wilt model with a precision of 0.98, a recall of 0.97, and an F1-score of 0.97, showing the high performance of the model in identifying the wilt symptoms with minimum classification errors. The F1-score of healthy leaves was also very high (0.96). Class results for the Cercospora, Anthracnose, and Mosaic Virus classes were also very good, demonstrating the efficacy of the proposed hybrid architecture in visually similar classes. The high weighted average precision, recall, and F1 score of 96.2%, 95.8%, and 96.0% further demonstrate the strength and generalization ability of the proposed framework for a wide range of categories of Capsicum leaf diseases.

V. CONCLUSION AND FUTURE WORK

In this study, we propose a novel hybrid CNN–ViT–CapsNet model tailored for dual domain recognition of Capsicum disease with both real-time and synthetic images. The proposed

framework has a unified framework that integrates local feature extraction, global contextual representation, and spatial validation for reliable disease classification across heterogeneous image sources, in contrast to the existing CNN-based or hybrid approaches. The architecture proposed combines the local feature extraction ability of CNNs, the global dependency modeling of Vision Transformers (ViT), and spatial hierarchy validation of Capsule Networks (CapsNet) into a single classification pipeline. It is a multi-branch fusion that creates robust disease diagnosis with orientation-invariant and heterogeneous image domain generalizability.

The proposed framework was extensively tested on a Capsicum leaf image dataset of 7,000 images and proved useful. Therefore, the proposed model shows classification accuracy of 96.8%, 97.5% AUC, while the baselines of CNN, ResNet50, VGG19, ViT, CapsNet, CNN-CapsNet and CNN-ViT have classification accuracy of 89.5%, 91.2%, 90.8%, 93.1%, 92.7%, 94.2% and 95.3%, respectively, and AUC of 90.2%, 92.1%, 91.5%, 94.0%, 93.3%, 94.8% and 95.6%, respectively. The framework achieved good results on both real-time (96.5%) and synthetic (97.1%) data sets, showing its effectiveness across the different image domains. Moreover, the model showed a high degree of flexibility with synthetic images generated using GANs, which helped to understand the performance level in a controlled dataset and its applicability to agricultural practices in the real world. The reliability of the framework was validated using performance metrics, as well as Precision, Recall, F1-score, and AUC, whereas the confusion matrix and ROC analysis indicated a significant increase in class separability, especially between diseases with similar symptoms like Cercospora and Anthracnose.

While good, there are some less-than-ideal aspects. The hybrid architecture requires more computation than the CNN-only architecture, and training requires a GPU to accelerate the training time. Moreover, although the amount of data used was enough to confirm the methodology, bigger multi-location data sets are required to test the scalability and flexibility of the framework even further.

A. Future Work

To extend this research, we plan to:

- Implement low-weight copies of the suggested model for edge computing and mobile applications, which will allow farmers to use the model in real-time in the field.
- Increase datasets with multi-spectral and hyperspectral images and capture physiological information that is out of the visible spectrum.
- Study self-supervised and transfer learning to minimise the need for big labelled data sets, especially in the case of rare Capsicum diseases.
- Investigate an end-to-end smart agriculture system with AIoT (AI of things) integration, including disease detection and farmer automated decision support.

Overall, the proposed CNN–ViT–CapsNet model represents a major leap forward in hybrid deep learning systems for agricultural disease detection, offering a scalable and comprehensive solution for precision farming technologies.

REFERENCES

[1] R. Kursun and M. Koklu, "ViT-AlexNet: A Hybrid Approach for Chili Leaf Disease Classification," Proc. Int. Conf. on AI and Applications, 2024. [Online]. Available: https://www.researchgate.net/publication/387823235

[2] K. Shaheed, I. Qureshi, F. Abbas, S. Jabbar, and Q. Abbas, "EfficientRMT-Net: An Efficient ResNet-50 and Vision Transformers Approach for Classifying Potato Plant Leaf Diseases," Sensors, vol. 23, no. 23, p. 9516, 2023. doi: 10.3390/s23239516

[3] T. Ali, B. S. Alqadi, M. Mehmood, and A. Nawaz, "CCNet: CNN-CapsNet-Based Hybrid Deep Learning Model for Diagnosing Plant Diseases Using Thermal Images," Proc. Int. Conf. on Computer Vision and AI, 2023. [Online]. Available: https://search.proquest.com/openview/464af0ece302e4a9f6cc518b1d44af7a

[4] I. Pacal, I. Kunduracioglu, M. H. Alma, and M. Deveci, "A Systematic Review of Deep Learning Techniques for Plant Diseases," Artificial Intelligence Review, Springer, 2024. doi: 10.1007/s10462-024-10944-7

[5] P. Dhaliwal, A. Srivastava, and S. Juneja, "Gabor-CapsNet and Hybrid ViT-CNN Models for Plant Disease Detection," in Handbook of AI in Agriculture, Springer, 2025. [Online]. Available: https://books.google.com/books?id=k8VoEQAAQBAJ

[6] P. Andal and M. Thangaraj, "Comprehensive Review of Methods for Leaf Disease Identification," Discover Artificial Intelligence, Springer, 2025. doi: 10.1007/s44163-025-00491-7

[7] V. Ahuja, A. Dogra, P. Kaushik, E. Jain, and V. Kukreja, "Vision Transformer-Capsule Network for Orange Quality Inspection: Ripeness and Black Spot Disease Detection," Applied Fruit Science, Springer, 2025. doi: 10.1007/s10341-025-01388-6

[8] M. Hussein and A. Mousa, "A Review of CNN-Based Techniques for Accurate Plant Disease Detection," Int. J. of Computers and Information, vol. 52, no. 3, pp. 233–244, 2023. [Online]. Available: https://journals.ekb.eg/article_317732.html

[9] K. R. P. Kumar and M. Gunasekar, "Fig Leaf Disease Prediction and Severity Estimation Using Deep Learning," in Proc. IEEE Int. Conf. on Computing and Applications, 2024. doi: 10.1109/ICCA.2024.10921856

[10] X. Zhang, Y. Mao, Q. Yang, and X. Zhang, "A Plant Leaf Disease Image Classification Method Integrating Capsule Network and Residual Network," IEEE Access, vol. 12, pp. 55621–55633, 2024. doi: 10.1109/ACCESS.2024.10472475

[11] A. Vinothini and R. V. Aswiga, "Transfer Learning Based Deep Learning Model for Classifying Tomato Plant Leaf Diseases," Engineering Research Express, vol. 7, no. 1, 2025. doi: 10.1088/2631-8695/add6f5

[12] L. Xu, J. Zhao, Z. Ma, J. Li, X. Wang, and Y. Liu, "A Review of Plant Leaf Disease Identification by Deep Learning Algorithms," Frontiers in Plant Science, vol. 16, 2025. doi: 10.3389/fpls.2025.1637241

[13] R. K. Naseer, A. Afzal, and M. Arshaq, "RiViT: Rice Leaf Disease Identification and Classification Using Vision Transformer (ViT)," in Proc. IEEE Int. Conf. on Machine Learning and Data Science, 2025. doi: 10.1109/MLDS.2025.11135614

[14] D. Zhang, B. Wang, and Z. Tian, "Stcovidnet: Automatic Detection Model Based on Swin Transformer," ResearchSquare, preprint, 2022. doi: 10.21203/rs.3.rs-1401026/v1

[15] S. Venkatraman, "A Channel Attention-Driven Hybrid CNN Framework for Paddy Leaf Disease Detection," arXiv preprint, arXiv:2407.11753, 2024. [Online]. Available: https://arxiv.org/abs/2407.11753

[16] R. Neware, "A Comprehensive Review on Detection of Horticulture Fruits Disease Using Machine Learning and Deep Learning Approaches," Preprints.org, 2025. doi: 10.20944/preprints202503.1769.v1

[17] A. Dolatabadian, T. X. Neik, and M. F. Danilevicz, "Image-Based Crop Disease Detection Using Machine Learning," Plant Pathology, vol. 74, no. 1, pp. 88–102, 2025. doi: 10.1111/ppa.14006

[18] H. N. Ngugi, A. E. Ezugwu, and A. A. Akinyelu, "Revolutionizing Crop Disease Detection with Computational Deep Learning: A Comprehensive

Review," Environmental Monitoring and Assessment, Springer, 2024. doi: 10.1007/s10661-024-12454-z

[19] S. Gupta, K. Rawat, K. Krishali, and N. Rawat, "A Transfer Learning-Based Approach for the Classification of Tomato Leaf Diseases Using Modified CNN Models," Int. J. of Computer Applications, vol. 185, no. 2, pp. 1–9, 2023. [Online]. Available: https://www.researchgate.net/publication/394792552

[20] M. P. Mathew, S. Elayidom, and K. M. Abubeker, "CapsNet Hybrid for Bell Pepper Leaf Disease Classification," ELCVIA, vol. 24, no. 2, pp. 33–42, 2025. [Online]. Available: https://elcvia.cvc.uab.cat/article/view/1970

Large-Scale Static Malware Detection Using Classical Machine Learning Models: An Evaluation on the EMBER Dataset

Achmad Fauzan[1], Tito Pinandita[2], Aulia Desy Nur Utomo[3]
Department of Informatics Engineering-Faculty of Engineering and Science, Universitas Muhammadiyah Purwokerto, Indonesia[1]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia[2]
Informatics Engineering Study Program, Telkom University, Purwokerto Campus, Indonesia[3]

Abstract*—Malware detection is a major difficulty in cybersecurity as malicious software continues to evolve in scale, diversity, and sophistication. While deep learning and highly complex architectures are becoming increasingly important in recent work, the practical efficiency of conventional machine learning methods for large-scale static malware detection remains underexplored. We perform a comparative evaluation of four machine learning models (Random Forest, XGBoost, Logistic Regression, and Decision Tree) on approximately 600,000 Portable Executable (PE) samples from the EMBER dataset. To enable a fair comparison of the models, we created a common experiment setup including standardised preprocessing, repeated evaluation with numerous random seeds, selective hyperparameter optimisation, feature importance analysis, and confusion matrix-based error analysis. The experimental results show a strong benefit of ensemble-based approaches for the structured feature representation provided by the EMBER dataset. Random Forest showed the best overall performance, with 96.74 % accuracy, 96.71 % F1-score, and a ROC-AUC of 0.9953, retaining a very steady behavior in repeated runs. XGBoost likewise demonstrated good predictive capacity with less training time but did not outperform Random Forest even with careful hyperparameter adjustment. On the other hand, Logistic Regression performed significantly worse, suggesting that linear decision boundaries were insufficient to capture the deep structural relationships encoded in static malware traits. Further study of the confusion matrix shows a balanced classification behavior with relatively low false negative rates, which is significant for operational malware detection situations. The feature importance analysis suggested that entropy-based features, PE structure metadata, and import-based features played an important role in the malware classification judgments. In conclusion, the results suggest that well-designed classical ensemble approaches are still quite competitive for scalable and interpretable static malware detection even with the rising usage of more and more powerful machine learning architectures.

Keywords*—Malware detection; static analysis; EMBER benchmark; ensemble learning; PE file analysis; random forest; XGBoost

I. Introduction

Malware detection has become increasingly difficult as modern cyberattacks continue to evolve in both scale and technical sophistication. In operational environments, security systems must process enormous volumes of executable files while simultaneously dealing with malware that frequently changes its structure, behavior, or observable characteristics to evade detection. Traditional signature-based approaches are still widely used because of their speed and practicality, yet their effectiveness declines substantially when facing polymorphic, packed, or obfuscated malware capable of modifying identifiable patterns with minimal effort. As a result, maintaining manually engineered signatures alone is no longer sufficient for addressing rapidly evolving threat landscapes. These limitations have contributed to the growing adoption of machine learning (ML) techniques that can identify broader detection patterns beyond previously observed malware variants [1], [2].

Among the various malware analysis strategies, static analysis remains one of the most practical solutions for large-scale detection pipelines. Unlike dynamic analysis, which requires executable files to run inside monitored environments, static analysis examines structural properties, metadata, and statistical characteristics without interacting with the runtime system. This approach significantly reduces computational overhead and enables efficient large-scale inspection of binary collections. However, static methods are not without limitations. Sophisticated malware can still employ obfuscation, encryption, or structural manipulation techniques specifically designed to conceal malicious indicators and reduce detection effectiveness [3], [4].

The development of machine learning-based malware detection has also been accelerated by the availability of large public benchmark datasets. One of the most widely adopted resources is the EMBER dataset, which provides structured feature representations extracted from Portable Executable (PE) files for reproducible large-scale experimentation [5], [6]. EMBER combines entropy-based measurements, PE metadata, import information, and hashed feature representations, allowing researchers to examine both statistical and structural characteristics of executable files within a consistent evaluation framework.

Recent studies increasingly emphasize deep learning architectures, including convolutional, recurrent, and transformer-based models capable of learning complex feature representations directly from raw or semi-structured malware data [7], [8], [9], [10], [11]. In parallel, growing attention has been directed toward adversarial robustness, transfer learning, and adaptive malware detection strategies designed to address

evolving threat landscapes [12], [13]. These developments demonstrate substantial progress in predictive capability, particularly in highly complex classification scenarios. Nevertheless, practical deployment considerations—including scalability, interpretability, computational feasibility, and reproducibility—remain important concerns in operational cybersecurity environments.

Classical ensemble learning methods continue to attract attention because of their ability to capture non-linear relationships while remaining comparatively stable in structured feature representations. Random Forest, in particular, has demonstrated strong robustness across a range of malware detection studies involving PE-based static features [14]. However, despite the growing preference for increasingly sophisticated architectures, the extent to which classical ensemble models remain competitive under large-scale static analysis conditions is still insufficiently understood.

Several limitations can also be identified within the current literature. Many existing studies rely on relatively small or imbalanced datasets, which may reduce the generalizability of reported findings. Comparative evaluations are often conducted under inconsistent preprocessing pipelines or heterogeneous experimental settings, making cross-study interpretation difficult. In addition, predictive performance is frequently discussed independently from computational behavior, feature contribution patterns, or classification error characteristics, even though these aspects are highly relevant for practical deployment scenarios [6], [7], [15].

This study addresses these gaps through a large-scale comparative evaluation of four machine learning models—Random Forest, XGBoost, Logistic Regression, and Decision Tree—using approximately 600,000 samples derived from the EMBER dataset. The study investigates whether classical ensemble methods remain competitive for large-scale static malware detection under highly structured feature conditions.

To investigate this premise, a consistent experimental framework was developed incorporating standardized preprocessing, repeated evaluation across multiple random seeds, selective hyperparameter optimization, feature importance analysis, and confusion-matrix-based error characterization. The study aims to examine how model complexity, feature structure, and computational behavior interact within large-scale static malware classification tasks. More broadly, the findings are intended to provide insight into the continuing practical relevance of classical machine learning approaches for scalable, interpretable, and operationally feasible cybersecurity systems.

II. Related Work

Machine learning has become one of the dominant approaches in modern malware detection, allowing security systems to identify malicious patterns that go beyond manually crafted signatures. Early work showed that classical algorithms such as Decision Trees, Random Forest, and Support Vector Machines can reliably separate malicious from benign binaries using structured feature representations [16], [17]. These studies, however, were typically conducted on dataset sizes several orders of magnitude smaller than the EMBER-scale collection used here, which limits how confidently their reported accuracy figures generalize to large, production-scale detection pipelines. As the field matured, attention shifted from raw classification accuracy toward feature quality, model adaptability, and long-term robustness across heterogeneous operating environments [18], [19] — questions that are addressed only partially in the present study but that motivate its emphasis on stability across repeated runs.

Ensemble-based methods have repeatedly outperformed simpler classifiers in malware classification, particularly on structured feature representations where the relationships between variables are strongly non-linear [14], [20]. Random Forest and related strategies tend to integrate heterogeneous features effectively while staying robust to noise, which makes them attractive for the scalable, computationally efficient pipelines that operational cybersecurity systems require [21], [22]. What is less consistent across this body of work is the preprocessing pipeline behind these results: some studies apply aggressive feature selection before training, while others, like the present one, retain the full processed feature set to preserve interpretability. This inconsistency makes raw performance numbers difficult to compare directly across papers, even when the underlying algorithm is the same — a point the present study tries to mitigate by holding preprocessing constant across all four models. Maintaining stable performance over time also remains an open problem, since malware continuously evolves through obfuscation, packing, and adaptive evasion [23], a limitation this study does not resolve but explicitly acknowledges in Section V.

Feature representation is central to static PE-based detection. Prior work has shown that entropy measurements, structural PE metadata, imported API functions, and string-based statistics carry a strong discriminative signal [4], [24], and the feature importance results reported here (Section IV-C) largely reinforce this pattern, with entropy- and import-based attributes again emerging as top contributors. EMBER extends these engineered features with large-scale hashed representations of imports and strings [5], [6]. While hashed features expand coverage, they come at the cost of interpretability, since individual hashed attributes cannot be inspected in isolation. The present study deliberately avoids this trade-off by retaining the complete processed feature representation for model training while reporting only a subset of interpretable features during feature importance analysis, which sacrifices some representational capacity in exchange for transparent feature attribution — a methodological choice that distinguishes it from studies that exploit the full hashed feature space for marginal accuracy gains.

A related line of work focuses on feature selection and dimensionality reduction to improve computational efficiency and reduce redundancy in large feature spaces [25], [26]. These approaches can shorten training time and, in some cases, improve stability, but identifying a universally optimal feature subset remains difficult because malware characteristics differ substantially across families and contexts [27], [28]. This is precisely why the present study takes the opposite approach: rather than searching for a reduced feature subset, it retains the complete interpretable feature set and instead controls variability through repeated evaluation across five random seeds

(Section III-E). The comparison suggests that feature-reduction strategies and seed-repetition strategies are addressing the same underlying problem — performance instability — from different angles, and a direct empirical comparison between the two remains an open question for future work.

Deep learning approaches occupy a different position in this landscape. Convolutional networks, hybrid architectures, and transformer-based systems have demonstrated strong predictive capability by learning hierarchical representations directly from raw or semi-structured malware data [7], [8], [10], and subsequent work has pushed further into transfer learning, representation learning, and adversarial robustness [9], [11], [12], [13], [29], [30]. These studies generally report higher peak accuracy than classical models, but rarely under conditions directly comparable to this study's evaluation protocol: many deep learning studies evaluate on smaller curated subsets, use GPU-dependent training pipelines, and do not report training time or statistical significance testing across repeated runs. This makes it difficult to determine whether reported accuracy gains stem from the architecture itself or from differences in dataset scale and preprocessing — a gap the present comparative framework is designed to close, at least within the classical-model space.

Model configuration also matters considerably for deep learning-based detectors. Training algorithm selection [31] and hidden-layer architecture design [32] can shift predictive performance substantially, but these gains typically come with added computational cost and optimization effort, raising practical concerns around deployment cost, scalability, and interpretability [21], [22], [33]. By contrast, the four classical models evaluated here were deliberately kept close to default configurations (with selective tuning applied only to XGBoost, as detailed in Section III-F), trading a degree of potential peak performance for a more transparent and reproducible comparison baseline.

Comparative evaluations across model families confirm that behavior depends heavily on both feature representation and dataset characteristics. Random Forest and other tree-based ensembles tend to capture non-linear relationships effectively while remaining stable under noise [26], [34], whereas linear models such as Logistic Regression often struggle in highly structured feature spaces, and single Decision Trees are prone to instability and overfitting [16], [17] — a pattern this study's own results (Section IV-A) reproduce almost exactly. Dataset composition compounds these effects, particularly when class imbalance, heterogeneous malware families, or shifting feature distributions are involved [15]. Because preprocessing pipelines and evaluation conditions vary so much from one study to the next, direct comparison across the literature is rarely straightforward; this is one reason the present study fixes a single preprocessing pipeline across all four models rather than optimizing each independently.

Within this landscape, EMBER has become one of the most widely used benchmarks for static malware detection [5], [6]. Existing EMBER-based studies have examined feature effectiveness, classification accuracy, temporal degradation, and concept drift under evolving malware conditions [6], [7], [35], [36]. A consistent finding across this work is that detection performance deteriorates once a model is deployed outside the temporal window of its training data [37], [38] — a concern this study does not directly test, since its train–test split is randomized rather than temporally ordered (a limitation discussed further in Section V). Other EMBER-based work has explored simplified, more interpretable feature representations and shown that reduced feature spaces can still achieve competitive detection accuracy [39], a finding broadly consistent with this study's own decision to retain the complete EMBER feature representation for model training while highlighting a subset of interpretable features during feature importance analysis.

Taken together, these studies report strong malware detection performance using both classical and deep learning approaches, but cross-study comparison remains genuinely difficult: dataset size, feature engineering choices, class balance, preprocessing steps, and evaluation protocols all vary simultaneously, making it hard to attribute reported performance differences to the learning algorithm alone. Few studies hold these factors constant while comparing multiple classical models side by side, and fewer still combine predictive accuracy with computational efficiency, feature importance analysis, and confusion-matrix-based error characterization within a single, unified experimental setup.

This study addresses that gap through a large-scale comparative evaluation of four classical machine learning models — Random Forest, XGBoost, Logistic Regression, and Decision Tree — on approximately 600,000 labeled EMBER samples, all processed through an identical preprocessing pipeline. Beyond classification accuracy, the evaluation framework incorporates repeated experiments across multiple random seeds, statistical significance testing, computational efficiency measurement, feature importance analysis, and detailed confusion-matrix characterization. This integrated design is intended to isolate algorithmic differences from the confounding methodological variation that complicates comparison across the broader literature reviewed above.

III. Materials and Methods

A. Dataset

The EMBER dataset was used as the primary data source for this study. EMBER provides pre-extracted static features derived from Portable Executable (PE) files and has become one of the most widely adopted benchmarks for machine learning-based malware detection research [5], [6]. The dataset contains both labeled and unlabeled samples, enabling large-scale evaluation under realistic malware analysis settings.

The EMBER 2018 training dataset contains 900,000 Portable Executable (PE) files, consisting of 300,000 benign samples, 300,000 malicious samples, and 300,000 unlabeled instances. Following the standard supervised learning setting adopted in the EMBER benchmark, unlabeled samples (label = −1) were excluded from the analysis. Consequently, the final dataset comprised 600,000 labeled PE files, including 300,000 benign and 300,000 malicious samples. This balanced class distribution reduces potential bias toward majority classes and enables fair comparison of machine learning models. The

resulting dataset was partitioned using a stratified 80:20 train–test split. Detailed dataset statistics are provided in Table I.

TABLE I. STATISTICAL SUMMARY OF THE DATASET

Property	Value
Total Samples	600000
Total Features	2381
Malicious Samples	300000
Benign Samples	300000
Train-Test Ratio	80:20
Number of Experimental Runs	5
Random Seeds	[42, 52, 62, 72, 82]
Chunk Size	5000

B. Data Preprocessing

Data preprocessing was performed to ensure consistency across all experimental stages and to improve compatibility with the evaluated machine learning models. Duplicate entries were removed to reduce redundancy, while samples labeled as −1, representing uncertain or unlabeled instances, were excluded from the analysis. Only samples labeled as 0 (benign) and 1 (malicious) were retained, following common practices in EMBER-based malware classification studies [36].

Feature scaling was subsequently applied using the StandardScaler implementation provided in the scikit-learn library. Numerical attributes were transformed into zero-mean and unit-variance distributions to reduce disparities in feature magnitude. This preprocessing step is particularly important for models such as Logistic Regression, where optimization behavior is sensitive to feature scale.

Standardization transforms each feature according to

$$z = (x - \mu) / \sigma \quad (1)$$

where, x denotes the original feature value, μ represents the feature mean, and σ denotes the standard deviation. This transformation improves numerical stability and reduces disparities in feature magnitude across variables.

Tree-based methods, including Random Forest and Decision Tree, are generally less affected by feature normalization. Nevertheless, all models were trained using the same processed dataset in order to preserve consistency throughout the comparative evaluation. Although scaling does not substantially influence tree-based learning behavior, a unified preprocessing pipeline was maintained to ensure experimental comparability across all classifiers.

C. Feature Processing

The EMBER dataset contains multiple categories of static Portable Executable (PE) features, including entropy measurements, PE header metadata, imported functions, string-based characteristics, and hashed representations extracted from executable binaries [5], [6]. In the present study, all numerical features generated during preprocessing were retained to preserve the full representational capacity of the processed dataset. Explicit dimensionality reduction or feature elimination techniques were not applied.

Missing values were handled during preprocessing to maintain dataset consistency before model training. The retained features included entropy-related measurements, PE header metadata, string-based representations, import-related characteristics, byte histograms, byte-entropy histograms, and other static Portable Executable attributes, all of which have previously been identified as informative indicators for malware classification tasks [4], [24].

After preprocessing, the resulting dataset contained 2,381 numerical features derived from the original EMBER feature representation. No feature selection, dimensionality reduction, or feature elimination procedures were applied. Consequently, all available processed features were retained and used during model training and evaluation.

For interpretability purposes, feature importance analysis focused on a subset of representative human-readable features originating from entropy-related attributes, PE metadata, imported functions, and string-based characteristics. These features were highlighted in the visual analysis because they provide intuitive explanations of model behavior, whereas many EMBER features are represented through hashed or aggregated feature encodings that are less directly interpretable.

D. Model Selection

Four machine learning models were selected to represent different learning paradigms and classification behaviors: Random Forest, Logistic Regression, Decision Tree, and Extreme Gradient Boosting (XGBoost). Random Forest, originally proposed by Breiman [40], was selected because of its strong performance on structured feature representations and its ability to capture complex nonlinear relationships through ensemble aggregation. Logistic Regression was included as a linear baseline to evaluate feature separability within the EMBER representation space. Decision Tree provides an interpretable non-linear alternative, while Extreme Gradient Boosting (XGBoost), introduced by Chen and Guestrin [41], was incorporated as a boosting-based ensemble framework capable of modeling complex feature interactions and serving as a strong contemporary benchmark.

More computationally intensive approaches, such as Support Vector Machines, were not included due to scalability limitations when applied to datasets of EMBER-scale dimensionality and size. The selected models, therefore, provide a balanced comparison across predictive performance, interpretability, computational behavior, and structural complexity [16], [20].

E. Training and Evaluation

The dataset was divided into training and testing subsets using a stratified 80:20 split configuration to preserve class balance across both partitions. This study adopted a randomized evaluation protocol because its primary objective was to perform a controlled comparative assessment of classical machine learning models under identical experimental conditions. Temporal validation based on malware collection chronology was not considered in the present work and remains an important direction for future research.

To reduce sampling bias and improve experimental robustness, all experiments were repeated across multiple random seeds. Performance results are therefore reported as mean values accompanied by standard deviations.

Classification effectiveness was evaluated using five commonly adopted metrics in malware detection research: accuracy, precision, recall, F1-score, and ROC-AUC [1], [15]. ROC-AUC was included to assess each model's ability to distinguish malicious and benign samples across varying decision thresholds, following the evaluation framework proposed by Fawcett [42].

The evaluation metrics were calculated based on the confusion matrix consisting of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy was computed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (2)$$

Precision as

$$Precision = \frac{TP}{TP+FP}, \quad (3)$$

Recall as

$$Recall = \frac{TP}{TP+FN}, \quad (4)$$

and F1-score as

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (5)$$

Performance variability across repeated runs was additionally analyzed to assess experimental stability and reproducibility.

To further investigate model behavior, confusion matrix analysis was conducted to examine false positive and false negative distributions. Feature importance analysis was also performed for tree-based models in order to identify the variables contributing most strongly to classification decisions.

The overall analytical workflow of the proposed methodology is illustrated in Fig. 1.

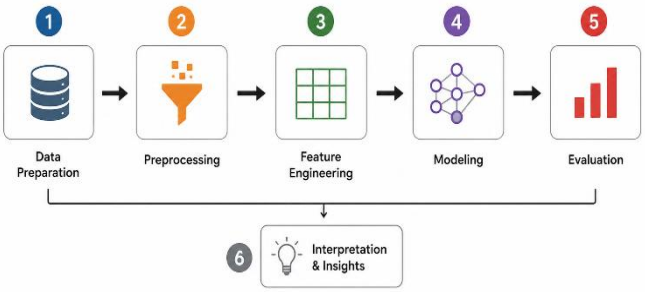


Fig. 1. Overall workflow of the proposed static malware detection framework.

F. Implementation Details

All experiments were conducted using Python 3.10. Machine learning models were implemented using the scikit-learn library (version 1.3) [43], while the XGBoost classifier was implemented using the XGBoost library (version 1.7).

The primary hyperparameter configurations used in the experiments are summarized in Table II. Most models were evaluated using default or near-default parameter settings in order to maintain consistency and reduce the risk of excessive configuration bias during comparison. Hyperparameter optimization was selectively applied to XGBoost using a randomized search strategy. The explored parameters included the number of estimators, learning rate, maximum tree depth, subsampling ratio, and regularization terms. Selective tuning was applied primarily to XGBoost because boosting-based ensemble methods are generally more sensitive to parameter configuration than bagging-based or linear approaches.

TABLE II. MAIN HYPERPARAMETERS OF THE EVALUATED MODELS

Model	Parameter	Setting
Random Forest	n_estimators	100
	max_depth	None
	min_samples_split	2
	random_state	Multiple seeds (42, 52, 62, 72, 82)
XGBoost	n_estimators	500
	max_depth	8
	learning_rate	0.05
	subsample	0.8
	colsample_bytree	0.8
	reg_lambda	1
	gamma	0
	random_state	Multiple seeds (42, 52, 62, 72, 82)
Logistic Regression	solver	lbfgs
	max_iter	1000
	random_state	Multiple seeds (42, 52, 62, 72, 82)
Decision Tree	max_depth	None
	min_samples_split	2
	random_state	Multiple seeds (42, 52, 62, 72, 82)

Models with lower sensitivity to hyperparameter variation were evaluated primarily under standardized configurations to preserve comparability and reduce unnecessary optimization bias across the experimental framework. Hyperparameter optimization was selectively applied to XGBoost using RandomizedSearchCV, as the predictive performance of gradient-boosting models is known to be highly sensitive to parameter configurations. The search process explored combinations of key parameters, including the number of estimators, maximum tree depth, learning rate, subsampling ratio, and column sampling ratio. The final configuration was selected based on cross-validation performance and subsequently used in all experiments. In contrast, Random Forest, Decision Tree, and Logistic Regression were evaluated using widely adopted baseline configurations to maintain methodological consistency and facilitate fair comparative analysis.

To reduce potential comparison bias, the objective of this study was not to identify the globally optimal configuration for

each classifier, but rather to evaluate representative implementations of commonly used machine learning algorithms under a consistent experimental framework. Random Forest, Decision Tree, and Logistic Regression were therefore evaluated using widely adopted baseline configurations, whereas XGBoost received limited hyperparameter optimization because its performance is known to be particularly sensitive to parameter settings. Consequently, the reported results should be interpreted as a comparison between a tuned boosting-based ensemble and standard implementations of alternative classifiers rather than as an exhaustive hyperparameter optimization study.

IV. Results

A. Performance Comparison of Machine Learning Models

The predictive performance of the evaluated machine learning models—Random Forest (RF), XGBoost, Logistic Regression (LR), and Decision Tree (DT)—was analyzed using five standard classification metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The corresponding results are presented in Table III.

Random Forest produced the strongest classification performance across nearly all evaluation metrics. The model achieved an accuracy of 0.9674 ± 0.0005, accompanied by a precision of 0.9769 ± 0.0007 and an F1-score of 0.9671 ± 0.0005. Its ROC-AUC value reached 0.9953 ± 0.0001, indicating an exceptionally strong capability to distinguish between benign and malicious samples across varying decision thresholds. The relatively small standard deviations observed across repeated runs further suggest stable behavior under different random initializations.

XGBoost (Tuned) also demonstrated high predictive capability, achieving an accuracy of 0.9493 ± 0.0007 and a ROC-AUC of 0.9906 ± 0.0002. Although the difference compared with Random Forest appears numerically modest, the separation remained consistent across all evaluation metrics. This indicates that both ensemble-based approaches adapted effectively to the engineered static PE feature representation, yet Random Forest maintained a slight but persistent advantage in classification consistency.

TABLE III. Performance Comparison of Machine Learning Models on the EMBER Dataset

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Training Time (s)
Random Forest	0.9674 ± 0.0005	0.9769 ± 0.0007	0.9574 ± 0.0007	0.9671 ± 0.0005	0.9953 ± 0.0001	48.5112 ± 0.5054
XGBoost (Tuned)	0.9493 ± 0.0007	0.9511 ± 0.0009	0.9474 ± 0.0006	0.9492 ± 0.0007	0.9906 ± 0.0002	20.3657 ± 0.3057
Logistic Regression	0.7402 ± 0.0009	0.7355 ± 0.0012	0.7502 ± 0.0012	0.7428 ± 0.0008	0.8047 ± 0.0010	4.3501 ± 1.2262
Decision Tree	0.9417 ± 0.0008	0.9385 ± 0.0015	0.9453 ± 0.0010	0.9419 ± 0.0008	0.9417 ± 0.0008	9.0032 ± 0.2118

To further evaluate whether the observed performance differences between Random Forest and XGBoost were statistically significant, paired t-test and McNemar test analyses were conducted. As shown in Table IV, both tests produced extremely small p-values, indicating that the observed differences are unlikely to be attributable to random variation alone. These findings provide additional evidence supporting the robustness of the comparative results obtained in this study.

TABLE IV. Statistical Significance Analysis Between Random Forest and XGBoost

Test	Comparison	P-value
Paired t-test	RF vs XGBoost	4.26×10^{-8}
McNemar Test	RF vs XGBoost	1.04×10^{-245}

Decision Tree achieved competitive results, with an accuracy of 0.9417 ± 0.0008 and a recall of 0.9453 ± 0.0010. The model showed reasonable discriminative capability despite its substantially simpler structure. However, its ROC-AUC value remained noticeably lower than that of the ensemble methods, suggesting reduced robustness when handling more ambiguous decision boundaries within the dataset.

A markedly different pattern emerged for Logistic Regression. The model obtained an accuracy of 0.7402 ± 0.0009 and ROC-AUC of 0.8047 ± 0.0010, substantially below the other evaluated approaches. Precision and recall also remained comparatively limited, indicating that linear decision boundaries were insufficient to fully capture the complex structural relationships embedded in the static PE features. The performance gap becomes particularly visible in Fig. 2, where the separation between linear and ensemble-based methods is clearly illustrated.

In terms of computational cost, Logistic Regression required the shortest training time, averaging 4.35 ± 1.23 s. Decision Tree and XGBoost followed with training times of 9.00 ± 0.21 s and 20.37 ± 0.31 s, respectively. Random Forest required the longest training duration at 48.51 ± 0.51 s. Even so, the increase in computational overhead was accompanied by consistently superior predictive performance, especially in ROC-AUC and precision, which are highly relevant in malware detection scenarios where false classifications may carry significant operational consequences.

A visual comparison of model performance is presented in Fig. 2.

The comparative results reveal a distinct separation between ensemble-based models and simpler linear approaches. Random Forest and XGBoost maintained consistently high scores across all metrics, suggesting strong compatibility with the statistical and structural characteristics of the EMBER feature representation. The stability observed across repeated experiments also indicates that the extracted features provide highly separable decision regions for tree-based learning methods.

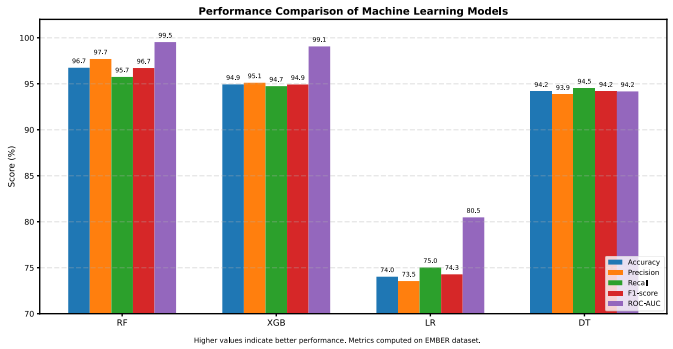


Fig. 2. Comparison of classification performance across evaluated machine learning models.

Another notable observation concerns the relationship between predictive performance and computational efficiency. XGBoost achieved substantially lower training time than Random Forest while still maintaining strong classification capability. Nevertheless, Random Forest consistently preserved higher accuracy, precision, and ROC-AUC values across all experimental runs. This behavior suggests that bagging-based ensemble strategies may provide stronger robustness than boosting approaches under highly structured static malware feature conditions.

To further evaluate whether the performance difference between Random Forest and XGBoost was statistically meaningful, paired statistical analysis was conducted. Both paired t-test and McNemar analysis indicated statistically significant differences between Random Forest and XGBoost ($p < 0.001$), confirming that the observed performance advantage of Random Forest was statistically reliable rather than attributable to random variation.

B. Confusion Matrix Analysis

To examine classification behavior in greater detail, confusion matrix analysis was performed for the Random Forest model. The corresponding results are presented in Fig. 3.

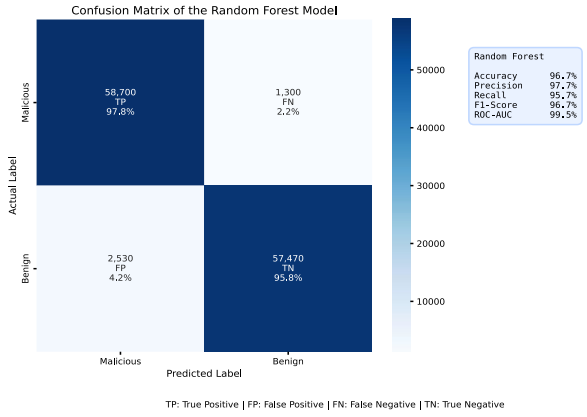


Fig. 3. Confusion matrix of the random forest model for malware detection.

The confusion matrix demonstrates that the model maintained a strong detection capability for both malware and benign samples.

A total of 58,700 malicious files were correctly identified as malware, representing 97.8% of the malicious class. At the same time, 57,470 benign samples were correctly classified, corresponding to 95.8% true negative recognition. These results are consistent with the high accuracy and ROC-AUC values reported previously in Table II, indicating that the model preserved stable discrimination across both classes rather than favoring a single category.

Misclassification rates remained comparatively limited, although several important patterns can still be observed. The model produced 1,300 false negatives, equivalent to 2.2% of malicious samples incorrectly predicted as benign. From a cybersecurity perspective, this category of error is particularly critical because undetected malware may continue to execute within operational environments without triggering defensive mechanisms.

The number of false positives was slightly higher, with 2,530 benign samples incorrectly classified as malicious, corresponding to 4.2% of the benign class. While false positives may increase operational overhead by generating unnecessary alerts or manual inspections, they are generally considered less severe than false negatives in malware detection systems, where missed threats can lead to direct security compromise.

Another notable characteristic of the confusion matrix is the relatively balanced classification behavior across both classes. The model did not exhibit substantial bias toward either malicious or benign predictions, suggesting that the ensemble structure of Random Forest was able to capture stable decision boundaries within the EMBER feature space. This balance is further reflected in the close alignment between precision (97.7%) and recall (95.7%), as illustrated in Fig. 3.

The distribution of errors also provides indirect insight into the nature of the EMBER dataset itself. Samples located near the decision boundary may contain overlapping structural or statistical characteristics, particularly in cases involving obfuscated binaries, packed executables, or benign software with behavior patterns resembling malicious activity. Despite these challenges, the relatively small proportion of misclassified samples indicates that the extracted static features remain highly informative for large-scale malware discrimination tasks.

C. Feature Importance Analysis

Feature importance analysis was conducted using the Random Forest model to identify the variables contributing most strongly to malware classification decisions. The ten highest-ranked features are presented in Fig. 4.

Among the evaluated attributes, strings.entropy emerged as the most influential feature, with an importance score of 0.1088. This result suggests that statistical irregularities within extracted string content provide highly discriminative signals for distinguishing malicious and benign executables. Elevated entropy values are commonly associated with packing, encryption, or obfuscation techniques frequently employed by malware to conceal internal behavior and evade static inspection.

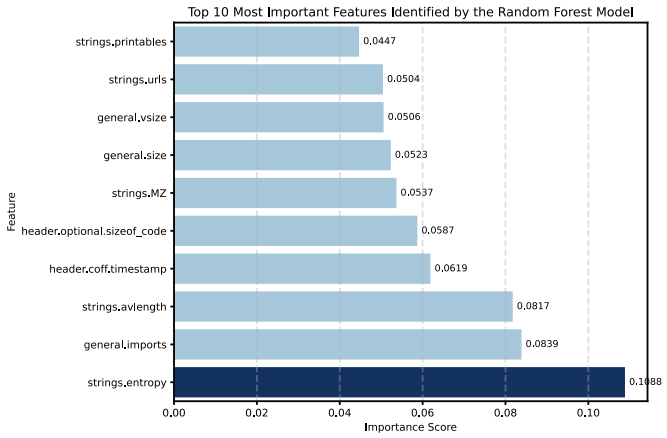


Fig. 4. Top 10 most important features identified by the random forest model.

The second and third most influential features were general.imports (0.0839) and strings.avlength (0.0817). The prominence of general.imports indicate that imported API patterns remain highly informative for malware identification, reflecting differences in functionality and execution behavior between benign and malicious software. Meanwhile, the importance of average string length suggests that textual artifacts embedded within executable files carry measurable structural information relevant to classification.

Several PE header and structural attributes also contributed substantially to model decisions. Features such as header.coff.timestamp (0.0619) and header.optional.sizeof_code (0.0587) ranked among the most relevant variables, indicating that metadata and binary organization continue to provide meaningful indicators of malicious activity. These attributes may capture indirect signatures related to compilation patterns, code organization, or manipulation of executable structures.

Additional features associated with binary size and embedded strings—including strings.MZ, general.size, general.vsize, strings.urls, and strings.printables—also demonstrated notable importance scores ranging from 0.0447 to 0.0537. Although individually less dominant than entropy-related attributes, their collective contribution highlights the multidimensional nature of the EMBER feature space, where both statistical and structural characteristics jointly influence classification behavior.

Another notable observation is the diversity of feature categories represented within the top-ranked variables. Rather than relying on a single dominant indicator, the Random Forest model leveraged complementary information derived from entropy measures, PE structural metadata, import-related characteristics, and string-based representations. This behavior likely contributed to the strong predictive stability observed in the previous evaluation results, particularly under repeated experiments across multiple random seeds.

The feature distribution shown in Fig. 4 also reinforces the suitability of tree-based ensemble methods for static malware analysis. Because Random Forest can capture complex non-linear interactions among heterogeneous attributes, the model is able to integrate multiple weakly correlated indicators into a robust classification framework. This characteristic is particularly advantageous in malware detection environments, where malicious samples often exhibit subtle variations designed to evade signature-based identification.

V. Discussion

The comparison between Random Forest and XGBoost is particularly noteworthy from a methodological perspective. Although XGBoost underwent selective hyperparameter optimization and achieved lower training time, it did not surpass the predictive performance of the bagging-based ensemble model. Furthermore, the superiority of Random Forest remained consistent across repeated experimental runs, suggesting that the observed differences cannot be attributed solely to parameter tuning. Instead, the findings indicate that the interaction between the learning mechanism and the structural characteristics of the EMBER feature representation may play a more influential role than algorithmic complexity alone. Similar observations regarding the effectiveness of ensemble tree methods in malware classification have been reported in previous studies involving structured PE-based feature representations [14], [26].

The results further suggest that feature structure may play a more decisive role than model complexity itself. The EMBER dataset contains a mixture of entropy-related attributes, import-based information, PE metadata, and hashed representations that collectively form highly informative decision regions. It should be noted that the reported feature importance rankings are limited to interpretable features presented for explanatory purposes. The predictive models themselves were trained using the complete set of 2,381 processed EMBER features. Ensemble methods appear particularly well-suited for integrating these heterogeneous feature categories while maintaining robust classification boundaries.

This behavior becomes more evident in the feature importance analysis. Entropy-based variables, especially strings.entropy, emerged as a dominant contributor to the classification process, indicating that statistical irregularities associated with packing, encryption, or obfuscation remain highly informative indicators of malicious activity. Import-related and structural PE features also contributed substantially, reinforcing the idea that executable organization and API dependency patterns preserve detectable traces even under code variation. These observations are consistent with prior work emphasizing the discriminative value of PE structural features in static malware analysis [4], [6].

A different pattern emerged for Logistic Regression. The model demonstrated substantially lower performance across all evaluation metrics, suggesting that linear decision boundaries are insufficient for capturing the complex relationships embedded within the EMBER feature space. Malware classification in large-scale static datasets appears to involve non-linear interactions among structural and statistical attributes, limiting the effectiveness of simpler linear models under realistic conditions.

The confusion matrix analysis provides additional practical insight into classification behavior. Random Forest maintained high true positive and true negative rates while preserving a relatively balanced error distribution across both classes. From

an operational cybersecurity perspective, the low false negative rate is particularly important because undetected malware may continue executing without triggering defensive mechanisms. Although false positives were still present, their proportion remained comparatively limited relative to the scale of the dataset.

The reported results should be interpreted within the context of a controlled benchmark evaluation. Because the objective of this study was comparative model assessment rather than deployment-oriented forecasting, model performance was evaluated using randomized train–test partitions rather than temporally separated malware collections.

Modern malware increasingly employs packing, encryption, and obfuscation techniques that can reduce the effectiveness of purely static feature representations. These conditions may partially explain the false negative cases observed in the present evaluation, where approximately 2.2% of malicious samples were incorrectly classified as benign.

From a computational standpoint, the findings also reveal an important trade-off between efficiency and predictive strength. Random Forest required longer training time than Logistic Regression and Decision Tree, yet the additional computational overhead remained reasonable considering the substantial improvement in detection capability. In practical deployment scenarios, such trade-offs are often acceptable when stronger malware discrimination significantly reduces the risk of missed threats.

Several limitations should nevertheless be acknowledged. First, the study relies exclusively on static analysis features extracted from Portable Executable files. While static analysis offers major advantages in scalability and processing speed, it may be less effective against sophisticated malware designed specifically to evade structural inspection through obfuscation or polymorphic transformation techniques [2].

Second, the evaluation was conducted using a temporally bounded snapshot of the EMBER dataset. Malware ecosystems evolve continuously, and feature distributions may shift over time as attackers adopt new evasion strategies, modify compilation behavior, or alter execution patterns across environments. Prior studies have shown that such concept drift conditions can significantly degrade malware detection performance when models are deployed beyond the temporal scope of their training data [37], [38]. As a result, continuous adaptation and periodic retraining mechanisms may become necessary to preserve long-term detection effectiveness.

Another limitation concerns the scope of the evaluated models. Deep learning architectures were intentionally excluded in order to focus on the balance between predictive performance, interpretability, and computational efficiency. However, representation-learning approaches may capture latent behavioral characteristics that are not fully represented through manually engineered static features. Similarly, the present evaluation did not explicitly measure operational deployment factors such as inference latency, memory consumption, or robustness against adversarial manipulation, all of which are becoming increasingly relevant in modern cybersecurity systems.

The evaluation was conducted on labeled samples only. Although this setting follows the standard EMBER benchmark protocol for supervised learning, future studies may investigate semi-supervised approaches that leverage the available unlabeled instances.

Furthermore, the evaluation employed a randomized train–test split rather than temporal validation. Consequently, the reported results reflect comparative model performance under controlled benchmark conditions and may not fully capture concept drift or evolving malware characteristics encountered in real-world deployment scenarios.

Another limitation concerns the potential overlap of malware families between the training and testing partitions. Although stratified random splitting is commonly adopted in EMBER-based studies, the present work did not explicitly enforce family-aware partitioning. Consequently, related malware families may appear in both subsets, potentially leading to optimistic performance estimates. Future research should investigate family-aware evaluation protocols to better assess model generalization to previously unseen malware families.

More broadly, the findings suggest that model selection in malware detection should be guided by the characteristics of the available feature representation rather than assumptions regarding algorithmic sophistication alone. Within the experimental setting considered in this study, classical ensemble methods provided a favorable balance between predictive performance, robustness, scalability, and interpretability.

VI. Conclusion

This study conducted a large-scale comparative evaluation of machine learning models for static malware detection using the EMBER dataset, with particular emphasis on predictive performance, computational efficiency, and classification behavior under consistent experimental conditions. The results demonstrate that strong malware detection capability can still be achieved through classical machine learning approaches when supported by informative and well-structured feature representations.

Among the evaluated models, Random Forest consistently produced the strongest predictive results across accuracy, precision, F1-score, and ROC-AUC metrics while maintaining stable behavior across multiple random seeds. The model also demonstrated balanced classification characteristics in the confusion matrix analysis, achieving high detection rates for both malicious and benign samples with relatively limited misclassification. Feature importance analysis further showed that entropy-based attributes, import-related patterns, and structural PE characteristics contributed substantially to classification decisions, indicating that static executable features continue to provide highly discriminative signals for large-scale malware analysis.

XGBoost also achieved strong classification performance, although its predictive capability remained slightly below that of Random Forest despite selective hyperparameter optimization. Within the scope of the two ensemble strategies compared here, this finding suggests that increasing boosting-model sophistication through tuning does not necessarily translate into measurable improvement once the feature space already exhibits

strong separability. In contrast, Logistic Regression showed substantially lower effectiveness, reinforcing the non-linear nature of malware classification within the EMBER representation space.

The findings also highlight an important practical implication for cybersecurity system design. Model selection should not be driven exclusively by trends toward increasingly complex architectures, but rather by the compatibility between learning mechanisms, feature structure, and operational requirements. In the present study, the bagging-based ensemble strategy provided a favorable balance between predictive robustness and computational feasibility without requiring excessive optimization procedures.

From an applied perspective, Random Forest remains a highly practical solution for scalable static malware detection environments. Its combination of strong predictive stability, interpretable feature behavior, and moderate computational requirements makes it suitable for deployment scenarios where reliability and efficiency are both critical considerations.

At the same time, several challenges remain open for future research. The current framework relies exclusively on static PE-based features and does not incorporate runtime behavioral analysis, temporal adaptation, or adversarial robustness evaluation. The evaluation is also limited to classical machine learning models; deep learning and transformer-based architectures, which have shown strong results in related EMBER studies, were not included in the present comparison and therefore cannot be directly assessed against the models evaluated here. Future studies may therefore benefit from integrating static and dynamic representations, benchmarking against deep learning-based detectors under matched experimental conditions, exploring adaptive learning strategies for evolving malware ecosystems, and evaluating resilience against adversarial manipulation in real-world operational settings.

Taken together, the results indicate that under highly structured feature conditions such as those provided by EMBER, a well-designed classical ensemble method like Random Forest achieved higher predictive performance than the tuned XGBoost configuration evaluated in this study. Whether this advantage extends to deep learning or other more complex architectures was outside the scope of this study and remains an open question for future comparative work.

ACKNOWLEDGMENT

The authors would like to thank the contributors of the EMBER dataset for making the dataset publicly available for malware detection research. During the preparation of this manuscript, the authors used ChatGPT (OpenAI) and QuillBot for language refinement and grammatical improvement. The authors reviewed and edited the generated content and take full responsibility for the content of this publication.

REFERENCES

[1] V. Rathod, C. Parekh, and D. Dholariya, “AI & ML-Based Anomaly Detection and Response Using Ember Dataset,” in 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, Sep. 2021, pp. 1–5. doi: 10.1109/ICRITO51393.2021.9596451.

[2] K. M. Balasubramanian et al., “Obfuscated Malware detection using Machine Learning models,” in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023, pp. 1–8. doi: 10.1109/ICCCNT56998.2023.10307598.

[3] A. Pathak, Th. S. Kumar, and U. Barman, “Static analysis framework for permission-based dataset generation and android malware detection using machine learning,” EURASIP J. Inf. Secur., vol. 2024, no. 1, p. 33, 2024, doi: 10.1186/s13635-024-00182-3.

[4] L. El Neel, A. Copiaco, W. Obaid, and H. Mukhtar, “Comparison of Feature Extraction and Classification Techniques of PE Malware,” in 2022 5th International Conference on Signal Processing and Information Security (ICSPIS), 2022, pp. 26–31. doi: 10.1109/ICSPIS57063.2022.10002693.

[5] H. S. Anderson and P. Roth, “EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models,” Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.04637

[6] Y. Oyama, T. Miyashita, and H. Kokubo, “Identifying Useful Features for Malware Detection in the Ember Dataset,” in 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), IEEE, Nov. 2019, pp. 360–366. doi: 10.1109/CANDARW.2019.00069.

[7] S. Pramanik and H. Teja, “EMBER - Analysis of Malware Dataset Using Convolutional Neural Networks,” in 2019 Third International Conference on Inventive Systems and Control (ICISC), IEEE, Jan. 2019, pp. 286–291. doi: 10.1109/ICISC44355.2019.9036424.

[8] G. Karat, J. Kannimoola, N. Nair, A. Vazhayil, S. V G, and P. Poornachandran, “CNN-LSTM Hybrid Model for Enhanced Malware Analysis and Detection,” Procedia Comput. Sci., vol. 233, pp. 492–503, May 2024, doi: 10.1016/j.procs.2024.03.239.

[9] A. Al-ojaimi, “Advanced Framework for Detecting Malware in Portable Executable (PE) Files Using a Multi-Model,” Journal of Information Systems Engineering and Management, vol. 10, pp. 769–781, May 2025, doi: 10.52783/jisem.v10i36s.6562.

[10] T. Kacem and S. Tossou, “Trandroid: An Android Mobile Threat Detection System Using Transformer Neural Networks,” Electronics (Basel)., vol. 14, no. 6, 2025, doi: 10.3390/electronics14061230.

[11] A. Tyagi, “Scaling deep learning models: Challenges and solutions for large-scale deployments,” World Journal of Advanced Engineering Technology and Sciences, vol. 16, pp. 10–20, May 2025, doi: 10.30574/wjaets.2025.16.2.1252.

[12] M. Eslamnejad, R. Taheri, M. Shojafar, and M. Bader-El-Den, “Federated learning-based robust android malware detection: label-flipping attacks and defenses,” Neural Comput. Appl., vol. 37, no. 32, pp. 27057–27082, 2025, doi: 10.1007/s00521-025-11656-x.

[13] D. Gibert, G. Zizzo, Q. Le, and J. Planes, “Adversarial Robustness of Deep Learning-Based Malware Detectors via (De)Randomized Smoothing,” IEEE Access, vol. 12, pp. 61152–61162, 2024, doi: 10.1109/ACCESS.2024.3392391.

[14] I. Darmawan, H. Setiaji, L. Sutriani, A. Supriyadi, R. Rizal, and A. Rahmatulloh, “Ensemble Learning for Malware Classification: A Performance Evaluation Using Random Forest and Hist Gradient Boosting,” May 2025, pp. 1–6. doi: 10.1109/ICoICT66265.2025.11193116.

[15] P. Mittal, H. S. Lallie, and E. Titis, “Impact of imbalanced datasets on ML algorithms for malware classification,” Information Security Journal: A Global Perspective, vol. 34, no. 3, pp. 251–264, 2025, doi: 10.1080/19393555.2025.2459736.

[16] R. Hasan et al., “Enhancing malware detection with feature selection and scaling techniques using machine learning models,” Sci. Rep., vol. 15, no. 1, p. 9122, 2025, doi: 10.1038/s41598-025-93447-x.

[17] A. M. Hilal et al., “Malware Detection Using Decision Tree Based SVM Classifier for IoT,” Computers, Materials and Continua, vol. 72, no. 1, pp. 713–726, 2022, doi: 10.32604/cmc.2022.024501.

[18] D. Dunsin, M. C. Ghanem, K. Ouazzane, and V. Vassilev, “Reinforcement learning for an efficient and effective malware investigation during cyber incident response,” High-Confidence Computing, vol. 5, no. 3, p. 100299, 2025, doi: 10.1016/j.hcc.2025.100299.

[19] D. Tomar, A. K. Verma, and K. Chhillar, “A Hybrid Static-Dynamic Malware Analysis Framework Using Interpretable Neural Network,”

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 09, May 2025, doi: 10.55041/IJSREM52505.

[20] Md. F. Bin Hafiz, N. A. Khan, Z. Kamal, S. Hossain, and S. Barman, "A Robust Malware Classification Approach Leveraging Explainable AI," in 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS), 2024, pp. 1–6. doi: 10.1109/ISCS61804.2024.10581382.

[21] Y. Hou, T. Truong-Huu, Z. Chen, C.-K. Kwoh, and S. G. Teo, "PROTEUS: Domain Adaptation for Dynamic Features in AI-assisted Windows Malware Detection," in 2023 IEEE International Conference on Data Mining Workshops (ICDMW), 2023, pp. 1322–1331. doi: 10.1109/ICDMW60847.2023.00170.

[22] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," IEEE Access, vol. 10, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[23] A. Sabbah, R. Jarrar, S. Zein, and D. Mohaisen, "Empirical Evaluation of Concept Drift in ML-Based Android Malware Detection," Jul. 2025. [Online]. Available: http://arxiv.org/abs/2507.22772

[24] M. Yousuf, I. Anwer, T. Shakir, M. Siddiqui, and M. Shahid, "A Multi-Feature Dataset for Windows PE Malware Classification," SSRN Electronic Journal, May 2023, doi: 10.2139/ssrn.4359519.

[25] T.-H. Lai, Y.-J. Tsai, and C.-L. Liu, "Improving the Performance of Static Malware Classification Using Deep Learning Models and Feature Reduction Strategies," Mathematics, vol. 13, no. 23, 2025, doi: 10.3390/math13233753.

[26] P. S. Nguyen, T. N. Huy, T. A. Tuan, P. D. Trung, and H. V. Long, "Hybrid feature extraction and integrated deep learning for cloud-based malware detection," Comput. Secur., vol. 150, p. 104233, 2025, doi: 10.1016/j.cose.2024.104233.

[27] H. Darabian et al., "A multiview learning method for malware threat hunting: windows, IoT and android as case studies," World Wide Web, vol. 23, no. 2, pp. 1241–1260, Mar. 2020, doi: 10.1007/s11280-019-00755-0.

[28] H. Mohammadian, G. Higgins, S. Ansong, R. Razavi-Far, and A. A. Ghorbani, "Explainable Malware Detection through Integrated Graph Reduction and Learning Techniques," Dec. 2024. [Online]. Available: http://arxiv.org/abs/2412.03634

[29] A. Bensaoud and J. Kalita, "CNN-LSTM and transfer learning models for malware classification based on opcodes and API calls," Knowl. Based. Syst., vol. 290, p. 111543, 2024, doi: 10.1016/j.knosys.2024.111543.

[30] Z. Zhao, S. Yang, and D. Zhao, "A New Framework for Visual Classification of Multi-Channel Malware Based on Transfer Learning," Applied Sciences, vol. 13, no. 4, 2023, doi: 10.3390/app13042484.

[31] H. Mustafidah, U. B. Pambudi, and S. Suwarsito, "Selection of the most accurate training algorithm in the backpropagation network based on the accuracy of data pattern matching," 2022, p. 060012. doi: 10.1063/5.0111243.

[32] H. Mustafidah, S. Geography, and S. N. C. Permatasari, "Accuracy of the Neurons Number in the Hidden Layer of the Levenberg-Marquardt Algorithm," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 4, pp. 2349–2353, Nov. 2019, doi: 10.35940/ijrte.D8259.118419.

[33] A. Agarwal, S. B. Verma, and S. Singh, "Real-Time Malware Prevention and Detection (MP&D) Framework in Cloud Computing Environments," in 2024 International Conference on Cybernation and Computation (CYBERCOM), IEEE, Nov. 2024, pp. 273–278. doi: 10.1109/CYBERCOM63683.2024.10803195.

[34] A. Alhogail and R. A. Alharbi, "Effective ML-Based Android Malware Detection and Categorization," Electronics (Basel)., vol. 14, no. 8, p. 1486, Apr. 2025, doi: 10.3390/electronics14081486.

[35] C. Galen and R. Steele, "Evaluating Performance Maintenance and Deterioration Over Time of Machine Learning-based Malware Detection Models on the EMBER PE Dataset," in 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, Dec. 2020, pp. 1–7. doi: 10.1109/SNAMS52053.2020.9336538.

[36] M. Şandor, R. M. Portase, and A. Coleşa, "Ember Feature Dataset Analysis For Malware Detection," in 2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, Oct. 2023, pp. 203–210. doi: 10.1109/ICCP60212.2023.10398693.

[37] F. Ceschin, M. Botacin, H. M. Gomes, F. Pinagé, L. S. Oliveira, and A. Grégio, "Fast & Furious: On the modelling of malware detection as an evolving data stream," Expert Syst. Appl., vol. 212, p. 118590, 2023, doi: 10.1016/j.eswa.2022.118590.

[38] A. Guerra-Manzanares, M. Luckner, and H. Bahsi, "Android malware concept drift using system calls: Detection, characterization and challenges," Expert Syst. Appl., vol. 206, p. 117200, 2022, doi: 10.1016/j.eswa.2022.117200.

[39] J. Mojžiš and M. Kenyeres, "Interpretable Rules with a Simplified Data Representation - a Case Study with the EMBER Dataset," in Data Analytics in System Engineering, R. Silhavy and P. Silhavy, Eds., Cham: Springer International Publishing, 2024, pp. 1–10. doi: 10.1007/978-3-031-53552-9_1.

[40] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[41] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[42] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[43] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, Jun. 2012.

A Data-Driven Visual Analytics Framework for Transaction-Level Retail Profit Modeling and Decision Support

Donia Badawood
Department of Data Science, College of Computing,
Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Retail companies are becoming increasingly dependent on data science to inform their pricing, assortment, and regional strategy decisions. Profitability drivers, however, are often difficult to interpret because they span product-level, geographic, discount, and operational environments. In most real-life applications, analytics and visualization are treated as two distinct processes, thereby restricting interpretability and undermining their ability to support decision-making. This study provides a decision-level visual analytics model of retail profit analysis at the transaction level. The model is illustrated on the publicly available Superstore dataset as a benchmark, which contains 9,994 order-line records between 2014 and 2017 with the variables time, product, geographic, customer-segment, shipping, sales, discount, and profit. The workflow combines feature engineering, hierarchical slicing, variance-based comparison, and five publication-ready dashboards covering temporal trends, product profitability, geographic heterogeneity, discount sensitivity, and fulfillment context. The overall profit margin in the analyzed dataset is 12.47%, and 18.72% of order lines are loss-making. The findings show uneven profitability across product groups and states, with loss pockets observed in sub-categories such as Tables and Bookcases, as well as in states like Texas and Ohio. There is also a distinct nonlinear discount behavior: profitability tends to be positive at discount rates below about 20%, then margins decline sharply, and loss rates increase exponentially at higher discount rates. These benchmark-specific results illustrate how integrated visual analytics can support structured inspection of pricing, portfolio, and region-related patterns within transaction-level retail data.**

Keywords*—Visual analytics; retail profitability; discount sensitivity; decision support systems; transaction-level analytics; retail data science; profit diagnostics; interactive dashboards

I. Introduction

The contemporary retail decision environment is characterized by an abundance of transactional data, yet constrained by limited analytical integration between modeling and visual interpretation. Transaction systems, e-commerce platforms, loyalty systems, and logistics networks generate extensive records of customer demand, performance data on fulfilling those demands, and commercial results [1]. However, it is often the managerial questions of most interest- where the profit is being made or not, which discounts are strategic or wasteful, how regional market structure is influencing performance- that cannot be answered with mere numbers.

The first challenge is that retail performance is multidimensional [2]. Even revenue growth cannot guarantee profitability, and product mix, discounting, logistics, and regional market conditions can undermine profits. The same commercial principle, even within the same retailer, may yield quite contrasting results across classes and regions [3]. Recent operations and pricing studies suggest that retailers typically face profit-waste or profit-service trade-offs in their inventory and perishable management, as well as in their markdown choices, particularly when confronted with heterogeneous demand and consumer behavior. Similarly, recent studies in the environment of production and operations indicate that it is possible to optimize markdown policies to overcome the processes of overstock and waste and protect profitability, as it demonstrates that it is not the strategy of a greater percentage of discount; it is a variable that should be controlled [4]. The organizational analytics literature, based on empirical evidence, indicates that the quality of information presentation in dashboard visualizations can affect decision quality through perceived task complexity and information satisfaction [2]. Similar research on dashboard adoption highlights that content and design may pose obstacles when dashboards impose excessive cognitive load, leading to lower acceptance despite a robust underlying data pipeline [3]. This means that retail analytics may fail not only in cases of poor case models, but also when information is not circulated and the organization cannot interpret, believe, and act on it.

Although retail analytics toolchains have become more mature, a structural gap persists: to date, in the academic literature and industry practice, analytical modeling and visualization are treated as disjointed, sequential steps rather than a cohesive decision-support system. In many implementations, predictive and diagnostic modeling is implemented in notebooks or analytics platforms, and dashboards are then compiled into a reporting layer. This split raises three recurring issues [5].

Firstly, the simplification of model results is a common practice aimed at concealing uncertainty, heterogeneity, and edge cases using a few measures or fixed graphs. Second, explainability is often introduced after the fact and not as a process of interrogation and counterfactual thinking. Third, assessment usually emphasizes model accuracy while underestimating decision performance, so the system will assist stakeholders in reaching superior conclusions more quickly and with fewer errors. The importance of an interactive visual

analytics system that combines computational intelligence with visualization to enhance the decision-maker's understanding is evident in recent work in visual decision support, where the underlying phenomena are complex and multi-source [6]. Similar studies in visual analytics and explainability suggest that visualization is not just a pretty product but a process of interpretability and trust, especially when a model's behavior is otherwise obscure [4]. Furthermore, systematic research on the assessment of visual analytics shows that analytical evaluation plans are required to ensure that visualization interfaces, in fact, facilitate analytical processes and insight generation, rather than merely enhance visual complexity [7]. The findings indicate that there is one obvious opportunity: retail analytics can be enhanced by closer integration between analytics and visualization, that is, explicitly decision-driven and assessable.

In this study, I address the above gap by introducing a decision-oriented visual analytics framework and demonstrating it using the publicly available Superstore benchmark dataset on Kaggle [8]. The benchmark is designed to illustrate the coordination of descriptive and diagnostic analytics for visualization in a transparent and reproducible way. This dataset is not intended to be representative of real retail operations, and the empirical results presented in this study are not meant to be retail-wide; they are seen as patterns unique to the benchmark.

To develop an end-to-end analytics workflow that identifies profitability drivers and discount-related risk patterns across product and regional dimensions.

To translate analytical findings into visualization constructs that support managerial decision tasks (diagnosis, prioritization, and policy adjustment).

To demonstrate how an integrated analytical and visualization approach improves interpretability and actionability compared with model-only or dashboard-only approaches.

From these objectives, I define the following research questions:

RQ1: How can transaction-level retail benchmark data be structured to support integrated descriptive and diagnostic visual analytics?

RQ2: Which dashboard structures best support inspection tasks such as identifying loss pockets, comparing product and geographic segments, and examining discount sensitivity?

RQ3: How does a coordinated multi-view workflow support interpretability relative to isolated summaries?

RQ4: What benchmark-specific patterns become visible when the proposed framework is applied to the Superstore dataset?

This work makes four distinct contributions to the fields of retail analytics and visual decision support:

A transparent decision-centric visual analytics workflow that links data preparation, derived indicators, and coordinated inspection tasks.

A structured dashboard architecture aligned with temporal, product, geographic, discount, and fulfillment-oriented diagnostic views.

An illustrative benchmark demonstration showing how heterogeneous patterns can be surfaced without treating the benchmark dataset as a basis for real-world retail generalization.

The rest of this study is organized as follows. Section II presents a literature review of related work in retail analytics, decision support, dashboards, and visual analytics. Section III outlines the Superstore dataset, feature engineering, and the formulation of the analytical problem, and provides the methodology, comprising preprocessing, descriptive and diagnostic analytics, and visualization design principles. Section IV presents findings that support profitability trends, discount-profit correlations, and regional heterogeneity, and provides a list of decision-oriented figures and tables. Section V presents the implications for retail decision-making and design recommendations for analytics dashboards. Section VI presents the conclusion and future work.

II. Related Work

The section situates the present research within a hybrid of three overlapping research directions: data science pipelines, retail analytics, and decision support visual analytics.

A. Data Science Pipelines for Business Analytics

Modern data science practice is structured as a pipeline that includes data acquisition and preparation, modeling and evaluation, and deployment to organizational decision systems. Data quality, feature engineering, and validation significantly affect reliability and business impact [9, 10]. In reality, however, technical pipelines can barely generate value. According to Schmidt [11], the failure of most analytics projects is not due to inaccurate models but to how they fit into business processes. Similarly, Mikalef et al. [12] indicate that the analytics potential must be aligned with organizational decision-making designs to provide measurable performance effects. These findings indicate that accuracy in modeling is not sufficient. More recent studies on decision support systems build on the work of analytics pipelines by incorporating interactive aspects. Basole et al. [13] state that AI-based analytics are expected to be deployed in visual analytics to support exploration and reasoning. Chakraborty and Rana [14] also demonstrate that the quality of the decision is determined by dashboard design and visualization. Explainable artificial intelligence is a research area that focuses on integrating interpretability mechanisms into workflows [15, 16]. This void is especially acute where retail data are heterogeneous, and only the decision-making cycle is fast.

B. Retail Analytics: Demand Patterns, Discount–Profit Relationships

The retail analytics literature spans demand prediction, price optimization, inventory management, and promotion design. One of the main points is the correlation between the demand stimulation (e.g., discounts, displays) and profitability. Li et al. [17] demonstrate nonlinear effects of discounting strategies on profit and waste in perishable retail, modifying the previously known understanding of discounting and its impact on profit.

Riesenegger and Hubner [4] show that markdown optimization is necessary to ensure that revenue is recovered and the margin is not compromised. The reaction of consumers to discounts is non-homogeneous and time- and presentation-sensitive [18, 19]. Profit outcomes in the omnichannel are further affected by regional demand variability and fulfillment costs [20]. Demand forecasting has been transformed into nonlinear forecasting, machine learning, and deep learning models that can now predict demand trends [21, 22]. Nevertheless, the accuracy of projections often does not provide the manager with insight. Retail managers must know why performance should differ across segments, rather than the approximate degree to which it is expected to differ. Recent literature indicates that there are still issues in converting analytical outputs into strategic retail decisions [23]. Besides, the aspect of spatial and segment heterogeneity is underdeveloped in integrated structures. Numerous studies report concentration findings or fixed results, and few offer evidence of the relationship between profitability dynamics at the Product, Region, and Customer group levels and operational decision-making.

C. Visualization and Visual Analytics for Decision-Making

Information visualization studies have determined that effective graphical representations require fewer cognitive resources and greater pattern recognition [24]. The idea is developed as visual analytics, which is computational analysis and interactive visualization to support arguments in complex fields [25]. The explainability and transparency of AI systems have been considered in recent work. The survey by Hohman et al. [15] elucidates explainable deep learning visual analytics, as visualization enables one to comprehend complex models. One review of visual analytics system evaluation techniques is Cui [26], which focuses on insight generation and out-of-the-box, usability-based measures of decision performance. Cognitive ergonomics and the use of the dashboard are required in organizations not only to enhance data quality but also to support cognitive ergonomics. Too much information can make them difficult to operate, as Allam demonstrates, due to cognitive load [27], and research has shown that visualization design can affect task trust and efficiency [14, 28]. The use of interactive capabilities such as filtering, drill-down, and brushing enables users to compare categories, regions, and time periods when analyzing retail data. However, even in real-world analytics, bad interaction design may complicate and make it less readable than it could be. This is why business intelligence, as a course of study, has brought to the fore the need to align visualization with decision-making initiatives and management activities within organizations. Despite this, predictive modeling and interactive visualization have not been fully achieved in most real-life applications. Explainable AI studies tend to focus on model-level explanations, whereas business intelligence studies focus on descriptive dashboards. The visualization design will need to be created not as a by-product but as an analytic companion to bridge these streams.

D. Synthesis: Limitations in Prior Work and Positioning of This Study

The reviewed literature reveals three structural gaps that collectively motivate the proposed framework, which focuses on reproducibility and performance and generally does not go beyond visualization as part of the analytical process. Model accuracy is prioritized, with secondary attention to decision interpretability. Second, existing retail analytics studies show that discounting and demand management are both economically complex, yet they usually present findings in aggregated or fixed-table forms. There are very few studies that specifically explore the role of interactive visualization in revealing heterogeneity in the discount-profit relation within product hierarchies and geographic areas. Third, the visual analytics research advances the ideals of interaction and cognitive support. The literature reviewed indicates a gap at the intersection of retail analytics, dashboard design, and visual analytics methodology. While a lot of effort has been devoted to forecasting, optimization, or static reporting, less effort has been devoted to compact, reproducible visual analytics workflows for transaction-level business benchmark data. In this work, we fill this gap by proposing a coordinated system for descriptive and diagnostic inspection and demonstrating its application to the Superstore benchmark dataset. The study demonstrates a level of analytical design and interpretability but does not yield validated retail policy inferences. To place the current research in the existing literature regarding retail analytics, data science pipelines, and visual analytics. Table I summarizes the main representative studies, emphasizing their analytical methods, data sources, visualizations, and limitations.

III. METHODOLOGY

The methodology is described as an analytical workflow that combines data preparation with descriptive and diagnostic analysis and decision-oriented visualization, rather than a strictly formal mathematical model. The methodology is designed to be reproducible, transparent, and explicitly oriented toward decision-focused retail analytics. Fig. 1 shows the workflow analysis of this study.

The process commences with the receipt of retail information (based on transactions), i.e., product hierarchy, geography, customer segments, and commercial variables (sales, profit, and discount). This data is then pre-processed and converted into predictive data, such as clean and verified data, along with suggested variables like profit margin, loss indicators, and shipping delays. Second step: performance-over-time data, aggregated master data, product-hierarchy profitability, geographic trends, and discount sensitivity are added to the data set. They are even more powerful data structures that enable descriptive and diagnostic analytics to identify profit concentration, pockets of loss, and regional performance variance. The outcomes of the analysis are delivered via a visual analytics dashboard, making it easy to observe the influence of time, product profitability, geographical heterogeneity, and price changes. Finally, the framework assists managers in making decisions by providing clear implications for discount governance, product portfolio optimization, and regional strategy refinement.

TABLE I. LITERATURE MATRIX SUMMARIZING PRIOR STUDIES ON DATA SCIENCE PIPELINES, RETAIL ANALYTICS, AND VISUALIZATION-BASED DECISION SUPPORT

Ref.	Study focus	Core methods	Data/domain	Visualization/ interaction	Evaluation style	Main takeaway	Key limitation (gap)
[4]	Markdown optimization under constraints	Optimization; policy modeling	Retail inventory	Static	Quantitative	Need to balance profitability vs other objectives	Weak integration with interpretability/visualization
[17]	Discounting + presentation impacts profit/waste	Analytical modeling; empirical evaluation	Retail (perishables)	Mostly static	Quantitative/field or model-based	Discount effects are nonlinear and context-dependent	Limited interactive exploration for the decision workflow
[18]	Price promotion response heterogeneity	Econometrics / causal inference (typical)	Retail transactions	Minimal	Empirical	Promo effects vary across customers/products	Often reported in aggregates
[19]	Promotion timing and depth	Empirical + modeling	Retail pricing	Minimal	Empirical	Timing/depth shapes margin outcomes	Limited decision-interface design
[20]	Omnichannel/regional cost and demand variation	Empirical/operations models	Retail supply + demand	Low–moderate	Applied empirical	Geography affects profitability via cost and service differences	Often lacks interactive geo drill-down
[21]	Retail demand forecasting (ML)	ML regression; feature engineering	Sales time series	Minimal	Predictive metrics	ML improves accuracy on nonlinear patterns	Doesn't ensure interpretability/actionability
[22]	Deep learning for forecasting	LSTM/TCN/transformers (typical)	Large retail time series	Minimal	Predictive benchmarks	Advanced models can outperform baselines	Often "black-box" to business users
[23]	Data-driven retail transformation	Conceptual + empirical synthesis	Retail organizations	Moderate	Case/field evidence	Adoption barriers include skills + usability	Limited concrete design for analytics-viz integration
[24]	Information visualization principles	Perception/cognition theory	Cross-domain	Core principles	Laboratory + theorys	Visual encoding affects comprehension	Doesn't prescribe domain-specific KPIs

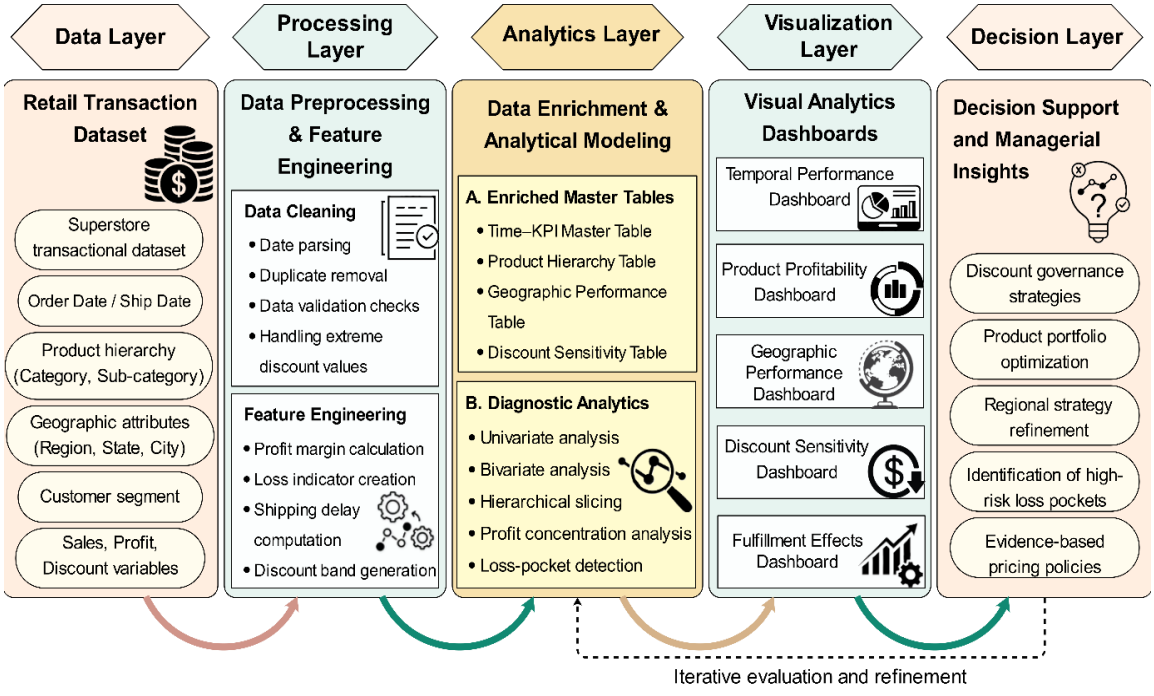


Fig. 1. Analysis pipeline of the proposed decision-centric retail data science framework.

A. Dataset Description

This study employs the publicly available Sample Superstore retail transactions dataset [8], having 9994 observations of order lines and 21 variables. Some records capture every product purchased in an order, including temporal, geographic, customer, product, and financial attributes. The database includes a couple of years of data, and data regarding the performance of transactions in four regions in the United States (South, West, Central, East), three groups of customers (Consumer, Corporate, Home Office), and three categories of upper-level products (Furniture, Office Supplies, Technology), which are subdivided into sub-categories. The temporal variables (Order Date, Ship Date), geographic variables (Region, State, City, Postal Code), segmentation of the customer (Segment), the product hierarchy (Category, Sub-Category, Product Name), operational variable (Ship Mode), and the commercial measures (s_i), Quantity (q_i), Discount (d_i), and Profit (π_i)) will be the major variables of each transaction.

Let the dataset be defined as:

$$\mathcal{D} = \{x_i\}_{i=1}^{9994} \tag{1}$$

where each observation is:

$$x_i = (t_i^o, t_i^s, g_i, c_i, seg_i, s_i, q_i, d_i, \pi_i) \tag{2}$$

This structure supports hierarchical aggregation, cross-sectional analysis, and time-based modeling. This study uses the publicly available Sample Superstore data as a benchmark to illustrate the proposed visual analytics structure. The dataset is popular in education and business intelligence presentations because it includes transaction-level retail information, including time, geographic, product, customer segment, discount, sales, and profit variables. The data used in the current research is not considered representative of the entire retail industry; it is used as a canned example to demonstrate how descriptive and diagnostic visual analytics can be integrated into a decision-support analytical process.

B. Data Schema and Variable Roles

I classify variables into identifiers, categorical descriptors, temporal features, and numerical values. The identifiers (Row ID, Order ID, Customer ID, and Product ID) are used for grouping and data integrity validation but are not considered explanatory predictors. For categorical variables, let:

$$\mathcal{C} = \{\text{Segment,Ship Mode,Region,State,City,Category,Sub-Category}\} \tag{3}$$

These variables define stratification levels for profitability and discount sensitivity analysis.

For the grouping variable G, define subset:

$$\mathcal{D}_g = \{x_i \in \mathcal{D} \mid G_i = g\} \tag{4}$$

For temporal variables, from Order Date t_i^o, I derive:

$$\phi_t(t_i^o) = \{\text{Year}_i, \text{Month}_i, \text{Quarter}_i\} \tag{5}$$

Shipping delay is defined as:

$$\Delta_i = t_i^s - t_i^o \tag{6}$$

This supports fulfillment performance and seasonality analysis.

For numerical variables, core commercial measures:

$$s_i \geq 0, q_i \in \mathbb{Z}^+, 0 \leq d_i \leq 0.8, \pi_i \in \mathbb{R} \tag{7}$$

Profit margin:

$$m_i = \frac{\pi_i}{s_i + \epsilon} \tag{8}$$

where, ϵ prevents division instability.

Aggregated group profit:

$$\Pi(g) = \sum_{x_i \in \mathcal{D}_g} \pi s_i \tag{9}$$

Aggregated group sales:

$$S(g) = \sum_{x_i \in \mathcal{D}_g} s_i \tag{10}$$

Group margin:

$$M(g) = \frac{\Pi(g)}{S(g) + \epsilon} \tag{11}$$

C. Problem Formulation

The analytical problem statement of the present research is as follows: where do profits lie, where do losses occur, and how can the patterns of losses and profitability be shown using hierarchical, temporal, geographic, and discount-based visual analytics. The study, rather than formal mathematical abstraction, emphasizes reproducible analytical steps that can aid interpretation and managerial diagnosis. The concept of a 'loss pocket' is defined for this study as a 'product group', 'geographic ', or 'discount group' that meets two criteria: 1) negative aggregate profit, and 2) materiality in terms of the visible aggregate contribution to the relevant 'slice of analysis '. As the benchmark study is descriptive in nature, materiality is treated as a threshold for interpretive salience, which is more of an operational definition. Negative profit and non-zero sales or order volume within the grouping are used to identify loss pockets in the present analysis.

1) Primary problem: Profitability Drivers and Discount Sensitivity: Retail managers should be familiar with what drives profitability and where discounting is killing the money.

I formalize profit as:

$$\pi_i = s_i - c_i^{eff} \tag{12}$$

where, c_i^{eff} represents the effective cost (implicitly embedded in the dataset's profit values).

I model profit drivers as:

$$\pi_i = f(d_i, q_i, seg_i, g_i, c_i, t_i) + \epsilon_i \tag{13}$$

The first objective is to identify group-level profit contributions:

$$\Pi(g, c) = \sum_{x_i \in (g,c)} \pi_i \tag{14}$$

The second objective is to quantify discount sensitivity:

Estimate conditional expectation:

$$\mathbb{E}[\pi \mid d] \tag{15}$$

Operationally, partition the discount into K bins:

$$B_k = \{i \mid d_i \in [a_k, a_{k+1})\} \tag{16}$$

Then:

$$\bar{\pi}_k = \frac{1}{|B_k|}\sum_{i \in B_k} \pi_i \tag{17}$$

The third one is to detect "loss pockets":

$$\text{Loss Pocket}(g,c) = \begin{cases} 1 & \text{if } \Pi(g,c) < 0 \text{ and } S(g,c) > \tau \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where, τ is a minimum materiality threshold.

The main issue that is to be resolved here is: 1) what sub-categories are value-destroying? 2) What are the levels of margin collapse at which the discount levels are? 3) Are there inefficiencies regarding regions in response to discounts?

2) Secondary problem: Profit Classification (Loss Risk Prediction): To support proactive decision-making, I define a binary classification task.

Define:

$$L_i = \begin{cases} 1 & \text{if } \pi_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

I aim to estimate:

$$P(L_i = 1 \mid X_i) \tag{20}$$

where the feature vector:

$$X_i = \{d_i, q_i, seg_i, g_i, c_i, \Delta_i, \phi_t(t_i^o)\} \tag{21}$$

A classifier f produces:

$$\hat{p}_i = f(X_i) \tag{22}$$

Decision rule:

$$\hat{L}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

This task identifies high-risk combinations of discount, category, and geography likely to result in negative profit. Although the classification of loss risks is conceptually applicable to the framework's future decision-support extension, the current research primarily focuses on descriptive and diagnostic analytics of the Superstore dataset. Thus, predictive modeling is considered a methodological extension rather than a key finding.

D. Data Preprocessing

Data preprocessing ensures structural consistency, numerical validity, and analytical readiness. Let the transactional dataset be defined as

$$\mathcal{D} = \{x_i\}_{i=1}^N \tag{24}$$

where each transaction x_i includes time, geographic, categorical, and financial attributes as defined in Section III.

1) Date parsing and temporal standardization: Order Date () and Ship Date () were parsed into standardized datetime formats. From I derived temporal features:$t_i^o t_i^s t_i^o$

$$\phi_t(t_i^o) = \{\text{Year}_i, \text{Month}_i, \text{Quarter}_i\} \tag{25}$$

Shipping delay was computed as:

$$\Delta_i = t_i^s - t_i^o \tag{26}$$

Transactions with $\Delta_i < 0$ were flagged as inconsistencies and verified. Negative delays were considered data-entry anomalies and excluded if confirmed erroneous.

2) Duplicate detection: Duplicates were examined using a composite key:

$$\mathcal{K}_i = (\text{OrderID}_i, \text{ProductID}_i, t_i^o) \tag{27}$$

If two records shared identical $\mathcal{K}_i$ and identical financial values, one instance was removed. Partial duplicates were retained when representing legitimate multi-line order structures.

3) Sanity checks and boundary validation: Financial and operational constraints were validated:

$$s_i \geq 0, q_i > 0, 0 \leq d_i \leq 1 \tag{28}$$

Profit values π_i were allowed to be negative, reflecting legitimate loss-making transactions. Extreme discounts (e.g., $d_i > 0.8$) were retained but flagged for sensitivity analysis due to their disproportionate impact on margins.

Profit margin was defined as:

$$m_i = \frac{\pi_i}{s_i + \epsilon} \tag{29}$$

where, ϵ prevents division by zero.

4) Feature engineering: To enable hierarchical and conditional analysis, additional features were constructed, i.e., 1) Profit margin, 2) Shipping delay, and 3) Binary loss indicator:$m_i \Delta_i$

$$L_i = \begin{cases} 1, & \pi_i < 0 \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

Categorical variables were encoded for modeling using one-hot encoding or target encoding, depending on the analytical stage.

E. Exploratory Data Analysis (EDA) Strategy

EDA followed a structured progression: univariate → bivariate → multivariate slicing.

1) Univariate analysis: For each numerical variable, summary statistics were computed:$X \in \{s, \pi, d, q, m\}$

$$\mu_X = \frac{1}{N}\sum_{i=1}^N X_i \tag{31}$$

$$\sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \mu_X)^2 \tag{32}$$

Distributional properties such as skewness were examined:

$$\text{Skew}(X) = \frac{1}{N}\sum\left(\frac{X_i-\mu_X}{\sigma_X}\right)^3 \quad (33)$$

This identifies heavy-tailed sales and profit distributions typical in retail.

2) Bivariate analysis: To evaluate discount–profit relationships, Pearson correlation was computed:

$$\rho_{d,\pi} = \frac{\sum(d_i-\mu_d)(\pi_i-\mu_\pi)}{\sqrt{\sum(d_i-\mu_d)^2}\sqrt{\sum(\pi_i-\mu_\pi)^2}} \quad (34)$$

Conditional profit means across discount bins B_k:

$$\bar{\pi}_k = \frac{1}{|B_k|}\sum_{i\in B_k}\pi_i \quad (35)$$

This provides a non-parametric view of discount sensitivity.

3) Segment and regional slicing: Group-level aggregation supports heterogeneity analysis. For the grouping variable :G

$$\Pi(g) = \sum_{i\in\mathcal{D}_g}\pi_i \quad (36)$$

$$S(g) = \sum_{i\in\mathcal{D}_g}s_i \quad (37)$$

$$M(g) = \frac{\Pi(g)}{S(g)+\epsilon} \quad (38)$$

Two-way slicing (e.g., Region × Sub-Category) was computed as:

$$\Pi(g,h) = \sum_{i\in\mathcal{D}_{g,h}}\pi_i \quad (39)$$

Variance decomposition assessed between-group effects:

$$\text{Var}(\pi) = \text{Var}(\mathbb{E}[\pi \mid G]) + \mathbb{E}[\text{Var}(\pi \mid G)] \quad (40)$$

The framework is a multi-hierarchical aggregate one, with the interpretation having to be sensitive to aggregation bias. Generalizations at the category, state, or regional level do not necessarily reflect behavior at the transaction level and vice versa. The framework thus promotes cross-level analysis of findings rather than limiting analysis to a summary table or graph. Sub-category and state-level inspection, in particular, is used as a safeguard against higher-level masking of the heterogeneity.

F. Descriptive and Diagnostic Analytics

In the present study, descriptive and diagnostic analytics are the focus. The analytical methods are designed to provide transparency in the inspection of profitability patterns over time, product hierarchy, geography, discount band, and fulfillment conditions. It is therefore more important to group aggregate data, provide comparative slicing, implement discount band analysis, and interpret cross-views than to model predictive analysis. Predictive techniques, such as loss-risk classification, would apply to potential future extensions of the framework but are not within the empirical focus of this manuscript.

G. Visualization Design Principles

In this context, visualization serves as a quantitative workflow companion, which is analytical in nature. Its visual design focuses on tracking time, hierarchical comparison, geographic heterogeneity, sensitivity to discounts, and operational context. It is intended to enhance interpretability and decision-making readiness, but not to maximize visual complexity.

$$\mathcal{D}' = \{x_i \in \mathcal{D}\text{:filter condition holds}\} \quad (41)$$

Drill-down supports hierarchical analysis, i.e., Region → State → Category → Sub-Category.

The uncertainty and variability for regression:

$$\hat{y}_i \pm 1.96\cdot\hat{\sigma} \quad (42)$$

For classification, the ROC curve is defined as:

$$\text{AUC} = \int_0^1 TPR(FP\,R^{-1}(u))du \quad (43)$$

Confidence intervals for group margins are computed using bootstrap resampling.

H. Evaluation Plan

This manuscript does not include descriptive/diagnostic benchmark demonstration. Does not report formal predictive evaluation and/or user-based decision-support experiments. Therefore, the framework is evaluated here for the transparency of the workflow, the coherence of the linked visual views, and the interpretability of the outputs, using a benchmark. The data were pre-processed by parsing the order and ship dates, deriving temporal fields, computing profit margin and shipping delay, screening for duplicates based on transaction identifier and matching financial data, as well as building grouped structures for category and state/region/discount-band inspection.

IV. Results

The results reported in this section are descriptive benchmark outputs. They are intended to support interpretive comparison across linked views and are not presented as statistically validated causal relationships. The analysis is conducted on the Superstore transactional dataset, comprising 9,994 order-line records spanning the period 2014 to 2017. Each order includes the order and shipping dates, customer group, product hierarchy (Category and Sub-Category), geographic information (Region, State, City), and core commercial results (Sales, Discount, Profit). To facilitate time-consistent reporting, I combined the data into time-based performance reporting, including sales, profit, margin, discount intensity, and the proportion of loss-making transactions. The time-highlighted summary is reported in the resulting time as shown in Table II.

Across all records, total sales are 2,297,201, and total profit is 286,397, corresponding to an overall profit margin of 12.47% (see Table II). Aggregate profitability is positive, but loss-making transactions are typical: 18.72% of order lines record negative profit. The healthy overall margin and high level of loss indicate an uneven distribution of profitability and an incentive for diagnostic breakdowns, further subsections (product hierarchy, geography, and discount sensitivity).

The distribution of the dataset is typical of a retail transaction: the sales distribution is right-skewed (there are many low-value order lines), and profit is left-skewed (there is a substantial tail), indicating a high loss rate. The concentration of discounts is low to moderate, and it includes discrete and recurrent components, which are standardized promotion

schedules rather than continuous discounting. All of the above properties indicate that aggregate statistics can mask clumps of concentrated negative profitability, which is why slice-based diagnostics by category, geography, and discount bands are needed in future result sections.

The multi-month performance view is shown in Fig. 2. The sales shown in Fig. 2(a) indicate seasonality and an upward trend over the study period. Monthly profit is depicted in Fig. 2(b), and its movement is sharper in relation to sales. The monthly profit margin is shown in Fig. 2(c), with periods of compressed margins and negative monthly margins. Fig. 2 (d) plots the monthly average discount and the monthly loss rate, and it is possible to visually check whether the greater the intensity of the discount, the greater the incidence of the loss-making transactions.

TABLE II. TIME-BASED KPIS ANNUAL AGGREGATION

Time period	Sales	Profit	Profit margin	Order lines	Avg. discount	Loss rate
2014	484,247.50	49,543.97	10.23%	1,993	0.1583	18.92%
2015	470,532.51	61,618.60	13.10%	2,102	0.1556	18.89%
2016	609,205.60	81,795.17	13.43%	2,587	0.1547	18.44%
2017	733,215.26	93,439.27	12.74%	3,312	0.1565	18.72%
Overall (2014–2017)	2,297,200.86	286,396.95	12.47%	9,994	0.1563	18.72%

[a] *Notes*: "Order lines" count the number of transactional records. "Loss rate" is the share of records with negative profit.

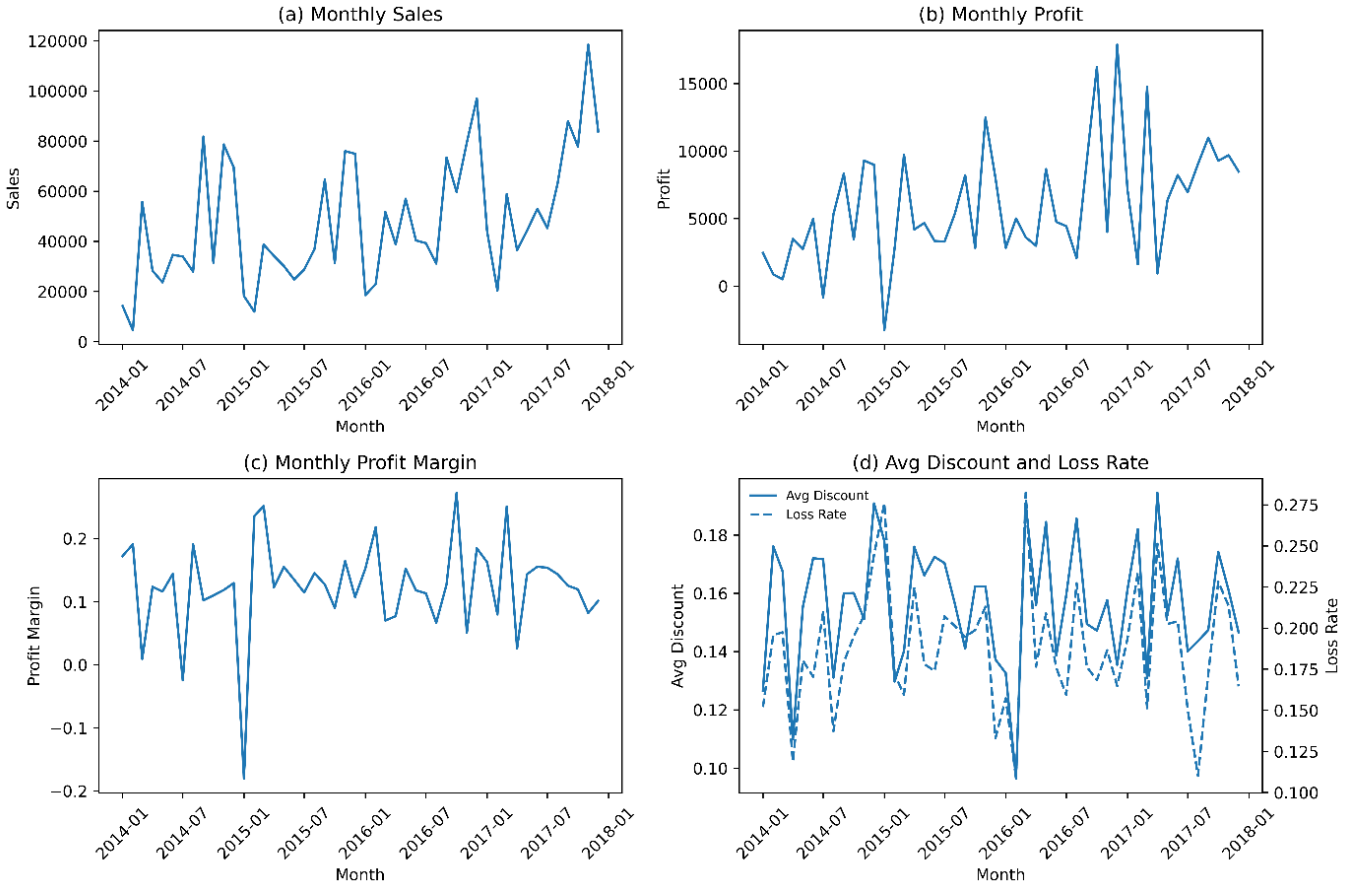


Fig. 2. Temporal performance dashboard.

The product hierarchy analysis indicates that structural disparities among the three top-level categories are high. As shown in Table III, technology records the highest total sales (836,154) and the highest profit contribution (145,455), which represents a margin of 17.40. Office Supplies is next with sales of 719,047 and profit of 122,491 (margin 17.04%).

Furniture, however, at sales of 741,999, yields only 18,451 in profit and a margin of 2.49, which is significantly low. This implies the existence of systemic profitability limits in the Furniture category. The sub-category analysis reveals a high level of profit concentration in Copiers, Phones, Accessories, and Paper. Copiers will generate a profit of 55,618 with a margin of 37.20 from only 68 transactions.

On the other hand, tables are the largest structural loss pocket in the hierarchy, with high sales volume, negative aggregate profit, and a loss rate exceeding 63%. Fig. 3 shows graphical summaries of these patterns: Fig. 3(a) compares category-level sales and profits side by side. These ranking patterns are shown in Fig. 3(b), and the spread of margins across subcategories is shown in Fig. 3(c). Fig. 3(d) shows the loss rates for sub-categories, with some lines consistently yielding negative results across all transactions.

Transaction-level heterogeneity is observed across subcategories. For example, Binders have negative volatility and high dispersion, yet show positive aggregate profit despite a high loss rate (40.25%). This type of internal dispersion makes it possible that aggregate category margins will obscure micro-level inefficiencies.

There is high geographic asymmetry. California brings the largest profit (76,381) from sales of 457,688 (margin 16.69%), and New York has the next-highest profit, 74,039 (margin 23.82%). The summary of geographic sales, profit, and margin distribution among the representative states is provided in Table IV, which reports differences in regional performance and identifies which states are structural loss pockets.

TABLE III. PROFITABILITY CONSTRAINTS WITHIN THE PRODUCT HIERARCHY

Category	Sub-Category	Sales	Profit	Margin	Avg. Discount	Orders	Loss Rate
Technology	Copiers	149,528	55,618	37.20%	0.162	68	0.00%
Technology	Phones	330,007	44,516	13.49%	0.155	889	15.30%
Technology	Accessories	167,380	41,937	25.05%	0.078	775	11.74%
Office Supplies	Paper	78,479	34,054	43.39%	0.075	1,370	0.00%
Office Supplies	Binders	203,413	30,222	14.86%	0.372	1,523	40.25%
Furniture	Tables	206,966	–17,725	–8.56%	0.261	319	63.64%
Furniture	Bookcases	114,880	–3,473	–3.02%	0.211	228	47.81%

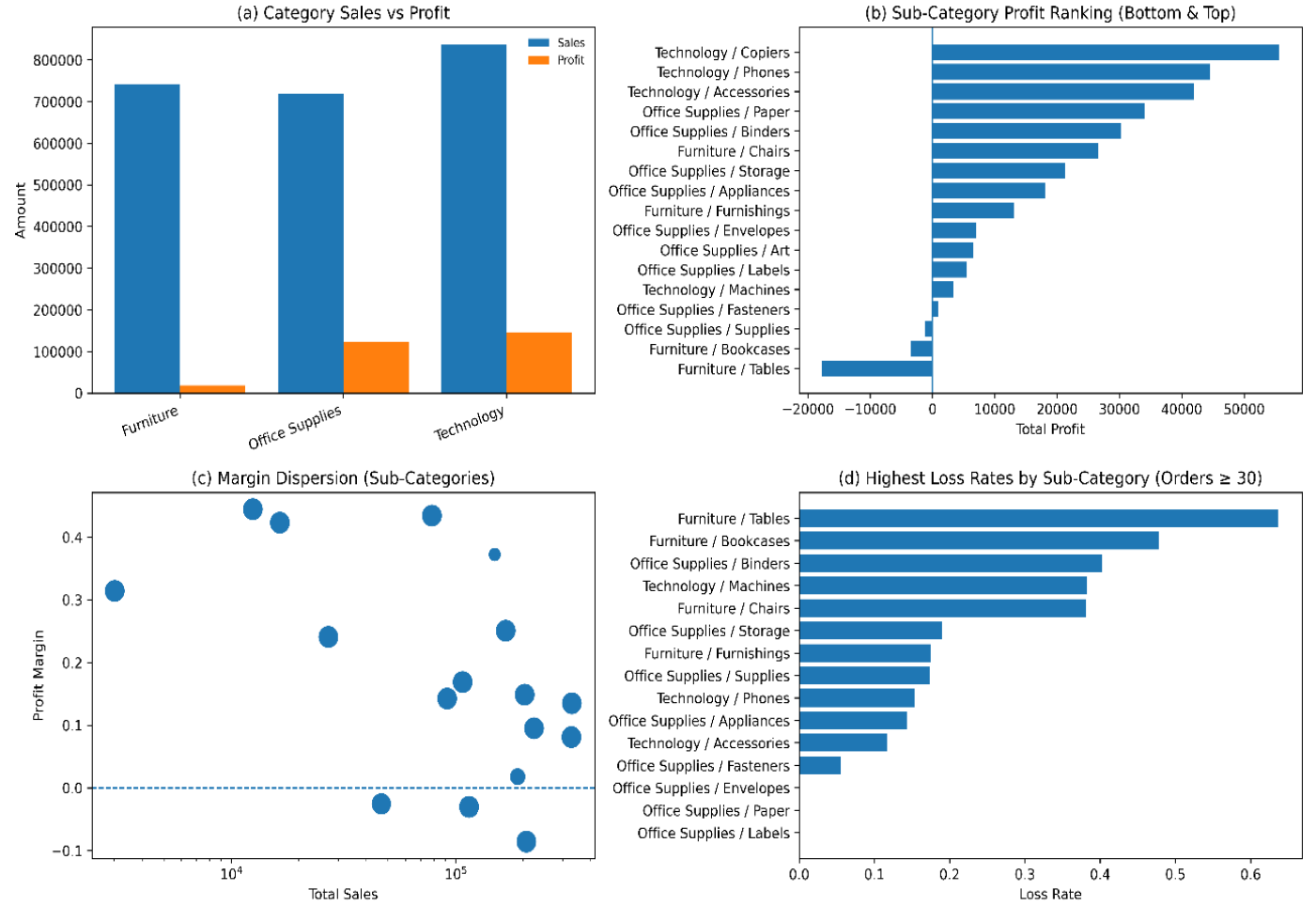


Fig. 3. Product profitability dashboard.

TABLE IV. GEOGRAPHIC PERFORMANCE (SELECTED STATES)

Region	State	Sales	Profit	Margin	Avg. Discount	Orders
West	California	457,688	76,381	16.69%	0.073	2,001
East	New York	310,876	74,039	23.82%	0.055	1,128
West	Washington	138,641	33,403	24.09%	0.064	506
Central	Texas	170,188	–25,729	–15.12%	0.370	985
East	Ohio	78,258	–16,971	–21.69%	0.325	469
Central	Illinois	80,166	–12,608	–15.73%	0.390	492

Texas, Ohio, Pennsylvania, and Illinois become significant pockets of losses, each with significant sales volume and negative aggregate profit. Other indicators of pricing-based margin erosion include these states offering higher-than-average discounts (averaging above 30%) and marking down margins. Central and Eastern states, by contrast, have negative persistent margins.

All these regional trends are summarized and visualized in Fig.4 (a). Fig. 4 (b) displays state-level dispersion, and Fig. 4 (c) depicts the sales-profit relationship among states. Cross-tabulation shows that Technology is running positive margins in the majority of regions, whereas Furniture performance exhibits considerable geographical variation. These interaction effects were illustrated in Fig.4 (d). States with negative margins are also high-risk, with high discount intensity. There is the greatest structural geographic loss pocket in Texas.

The mean discount across all transactions is 15.63. There is, however, an uneven distribution of discounts, and higher-value discount bands are associated with negative margins. The correlation between discount intensity and profitability is outlined in Table V, which reports sales, profit, margin, order volume, and loss rate by discount band. As shown in the table, the threshold effect is that profitability declines rapidly when discounts exceed 20%.

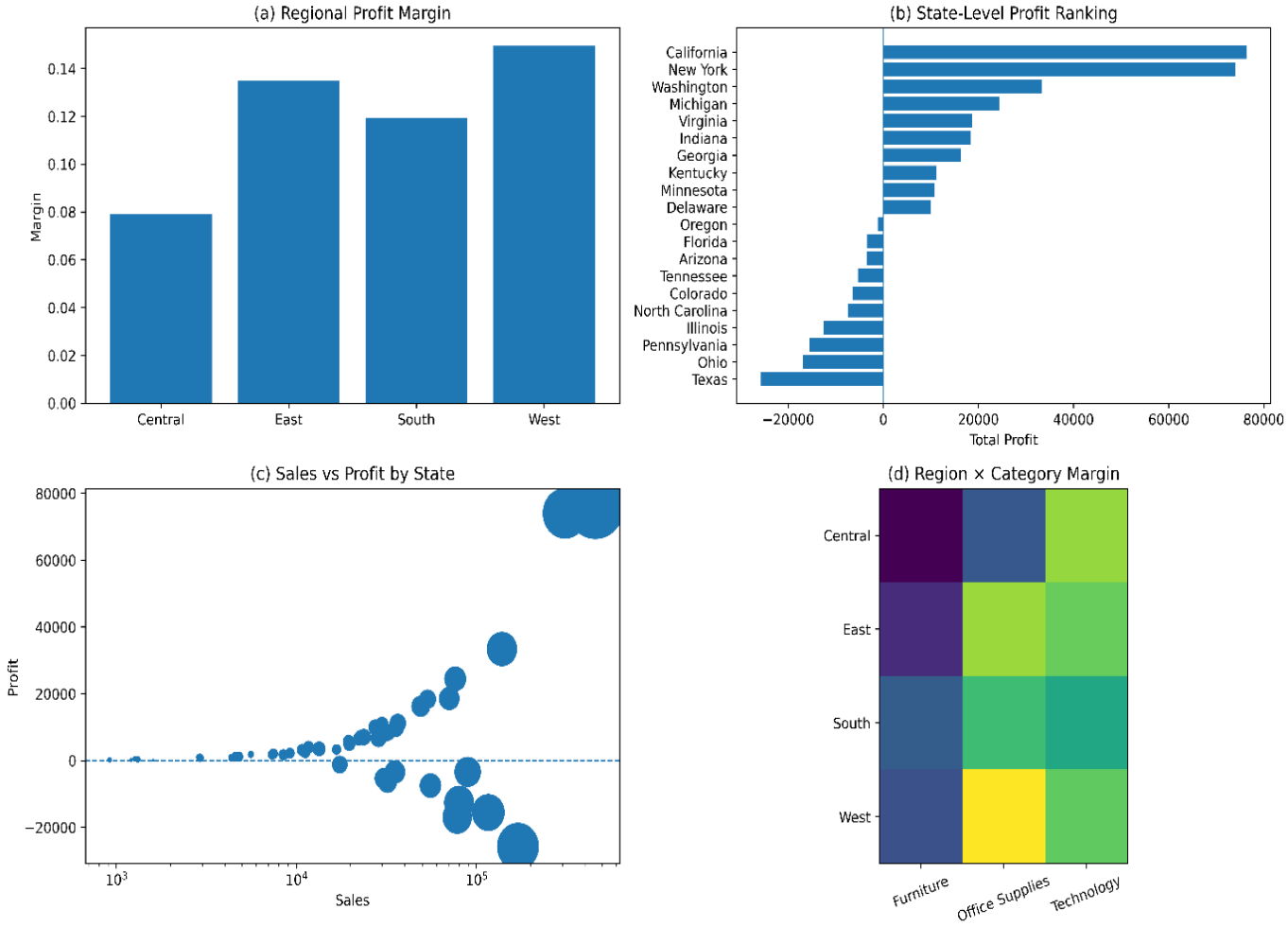


Fig. 4. Geographic performance dashboard.

TABLE V. DISCOUNT SENSITIVITY ANALYSIS

Discount Band	Sales	Profit	Margin	Orders	Loss Rate
0–10%	54,369	9,029	16.61%	94	4.26%
10–20%	792,153	91,756	11.58%	3,709	13.99%
20–30%	103,227	–10,369	–10.05%	227	91.63%
30–40%	130,911	–25,448	–19.44%	233	88.84%
40–50%	64,404	–22,999	–35.71%	77	100%
50%+	64,229	–76,559	–119.20%	856	100%

Table V summarizes only transactions included in the defined discount-band grouping used to analyze discount sensitivity. Therefore, the values in this table are not the totals of the entire dataset in Table II. Instead of replicating the aggregate totals for all transactions, the table shows comparative behavior across discount periods. The Discount Sensitivity Dashboard is shown in Fig. 5. Fig. 5(a) shows the number of orders by discount band, with the majority of transactions in the 10-20% and 50%+ bands. Fig. 5(b) shows profit by discount band, with positive profit at low discount rates and negative profit at higher discount rates. Fig. 5(c) shows the profit margin by discount band, indicating an apparent decline beyond the 20% discount mark, with very negative values beyond the 50% mark. Fig. 5(d) shows the loss rate by discount band, indicating that loss-making transactions are more than 20% off and 100% in the 40-50% and 50%+ bands, respectively.

Standard class accounts for the majority of transactions (5,968 orders) with an average delay of approximately 5 days. Same-day shipping exhibits minimal delay (0.04 days). Shipping delay alone does not exhibit a strong structural linkage to margin collapse. The discount intensity, rather than fulfillment speed, is closely related to loss behavior. Section V will discuss the research question of whether operational factors, i.e., shipping mode and delivery delay, are the main contributors to profitability variance. Although Fig.6 (a) shows significant differences in delivery speed among ship modes, Fig.6 (b) and Fig.6 (d) indicate that profit margins and loss rates do not differ significantly across fulfillment variants. In addition, as Fig.6 (c) shows, there is no apparent monotonic correlation between the delay length and transaction-level profit. A combination of these results indicates that the concept of operational efficiency is relevant to service quality but is not structurally descriptive of margin erosion in this dataset.

Table VI outlines the empirical findings of the proposed framework and representative studies in retail analytics and pricing research, and reveals that the proposed framework identifies multiple discount sensitivities, product-level diagnostics, and geographic profitability analyses differently.

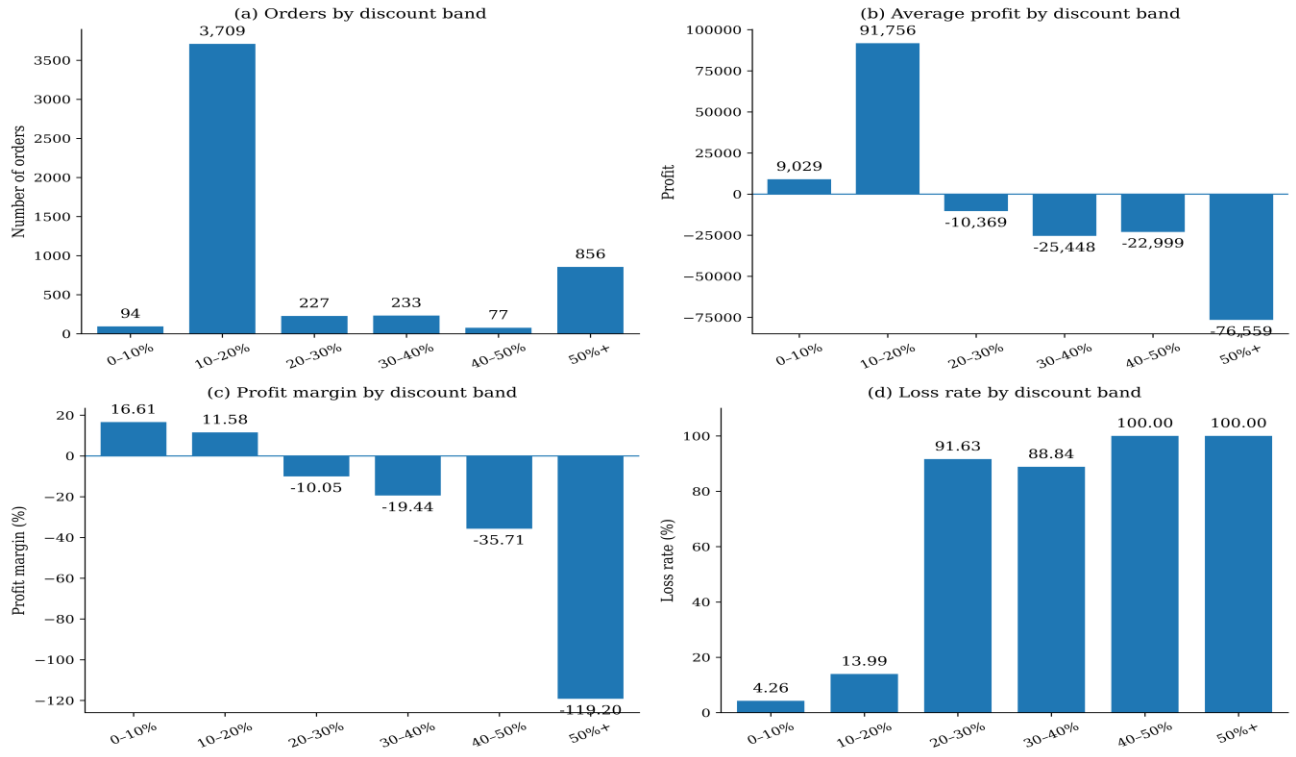


Fig. 5. Discount sensitivity dashboard.

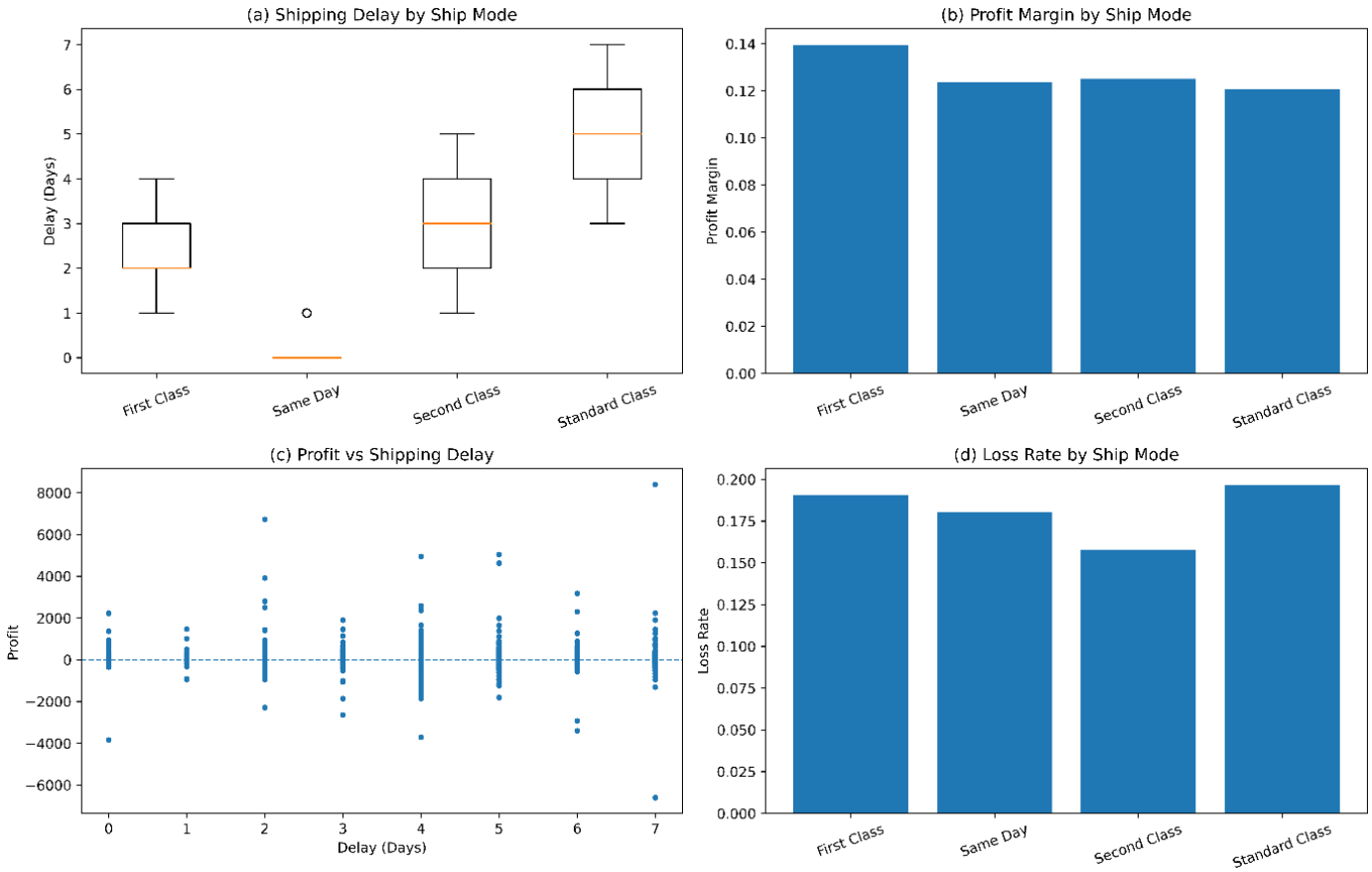


Fig. 6. Fulfillment effects dashboard.

TABLE VI. COMPARATIVE ANALYSIS OF RETAIL PROFITABILITY AND DISCOUNT SENSITIVITY FINDINGS

Ref.	Profitability Pattern	Discount Threshold Effect	Product-Level Variability	Geographic Impact	Visualization Support	Evaluation Type	Domain
[18]	Profitability is affected by discount and product presentation	Nonlinear discount effects reported	Product-dependent outcomes	Limited geographic analysis	× Mostly static analysis	Analytical/empirical modeling	Retail perishables
[19]	Profit depends on balancing markdown policies and inventory constraints	Discount policies influence profitability	Product-level variation acknowledged	Limited geographic dimension	× Static optimization models	Optimization analysis	Retail inventory
[20]	Promotion effects vary across products and customers	Promotion depth affects sales response	Strong product heterogeneity observed	Not central focus	× Minimal visualization	Econometric empirical analysis	Retail transactions
[21]	Profit affected by logistics and regional cost differences	Discount not primary focus	Product mix influences profitability	✓ Regional demand and cost differences	Limited	Empirical operations analysis	Retail supply chain
[25]	Visualization improves understanding of complex datasets	Not focused on pricing	Not domain specific	Not domain specific	✓ Visualization principles	Experimental evaluation	Cross-domain analytics
Proposed Framework	Overall margin 12.47%, loss rate 18.72%	Strong threshold near 20% discount; loss rate rises to 91–100% beyond threshold	High variability across sub-categories (e.g., Copiers high profit, Tables negative margin)	Strong regional disparity (e.g., Texas and Ohio negative margins)	✓ Multi-panel visual dashboards	Empirical transaction analysis	Retail transactional data

V. DISCUSSION

This section interprets the empirical findings in relation to the research questions, derives theoretical implications for integrating analytics and visualization, and outlines actionable managerial recommendations. Another practical issue concerns trade-offs among managerial goals. In the real retail world, an activity that boosts sales may compromise margin, and an activity that maintains margin may limit market growth or promotion. The proposed framework isn't a solution to these trade-offs itself, but is intended to put them into the spotlight. Positioning sales, profit, margin, discount, and geographic perspectives side by side provides a framework for discussing other goals and objectives that may compete for the solution, rather than assuming they are all equally important in every aspect of decision-making.

A. Answers to the Research Questions

RQ1: The benchmark analysis shows that profitability is unevenly distributed across product categories, sub-categories, and geographic units. Does the coordinated framework help reveal these benchmark-specific differences through linked hierarchical and geographic views?

The results indicate that profitability is not shared equally in product categories, subcategories, and geographic units. The category analysis reveals that Technology and Office Supplies have a margin of more than 17, whereas Furniture has a structurally weak margin of approximately 2.5. Sub-category profitability concentration is enormous: Copiers and Paper are making disproportionate margins, whilst Tables and Bookcases are making steady margin losses. These findings indicate that product mix and category structure are the most important in aggregate profitability.

RQ2: The discount-band inspection suggests a benchmark-specific threshold-like pattern in which profitability remains positive at lower discount levels but becomes negative in higher bands. This pattern is descriptive and serves as an example of the type of relationship that the framework can surface for further scrutiny.

The discount sensitivity analysis suggests a nonlinear threshold effect: profitability is generally maintained at discount levels below about 20%, whereas higher discount levels are associated with negative margins and sharply increased loss rates. The loss rates increase with higher discount levels and reach saturation at extreme levels. The findings of the present research confirm that discounting is not only positively related to low profitability but also a structural factor in margin erosion and the development of a loss pocket.

RQ3: The geographic views indicate that benchmark profitability varies across states and regions. The framework is useful here because it allows analysts to compare scale, margin, and product-region combinations across multiple linked views rather than relying on a single map or table.

The geographic test shows that profitability levels vary greatly across states. The other states, such as California and New York, that have been doing well show high margins, but other states, such as Texas and Ohio, have been incurring consistent losses despite high sales volume. The differences are associated with high discount intensity and the product mix structure, implying that the geographic strategy has a strong impact on financial performance.

RQ4: The fulfillment view suggests that shipping mode and delivery delay are less visually salient in this benchmark than discount and product structure. The framework, therefore, helps analysts prioritize which dimensions appear most informative within the benchmark context?

The fulfillment analysis demonstrates that shipping mode and delivery delay exert only a marginal influence on profit variability. Although Standard Class shipping involves longer delays (approximately 5 days), the profit margins and loss rates across shipping modes do not differ substantially. No monotonic relationship between delivery delay and transaction-level profit was observed. These findings indicate that while operational efficiency remains relevant for service quality, it does not structurally account for margin erosion in this dataset. The primary profitability drivers are pricing decisions and product mix effects, rather than logistics performance.

Taken together, the analysis of the results reveals that pricing choices, specifically discount policy, with product-hierarchy effects and geographic concentration, are used to explain most of the variation in profitability in the data.

B. Theoretical Implications: Integrating Analytics and Visualization

This study contributes to the convergence of three research streams: data science pipelines, business analytics, and visual analytics for decision support, showing that descriptive statistics alone cannot be used to identify structural risk. The aggregate profitability masks concentrated areas of loss that can be identified only through hierarchical slicing and multi-level aggregation. Second, the interpretation of the visualization and the analytical modeling is more interpretive. The multi-panel dashboards (e.g., discount sensitivity and geographic heterogeneity) show nonlinear threshold behavior and interaction effects, which would otherwise be difficult to see in tabular summaries alone. This visualization is not a reporting aspect, but an analytical tool that demonstrates structural relationships. Third, the granular analysis of retail transactions and the aggregation tables show that both may be applicable in retail. The profitability drivers all work in parallel at the micro (transactions), meso (sub-category/state), and macro (category/region) levels. The circulation across levels is consistent with a single, coherent pipeline spanning preprocessing, analytics, and visualization. In these respects, this work contributes to a theoretical framework in which analytics and visualization are not two separate components of data science systems based on decisions, but rather both are components of the system.

C. Illustrative Decision Scenarios Enabled by the Framework

As the dataset is a benchmark rather than a validated operational dataset, the framework should be used as an illustration of decision scenarios, not as a prescription. For instance, if there were a threshold-like discount pattern in a real deployment, the analysts might consider whether specific discount bands are correlated with weak margins in certain categories or regions. Likewise, regions with negative profits

may spur diagnostic follow-up, but not necessarily policy action. It is not a good framework to use if it is merely a set of predetermined managerial responses; it is a good framework to use because it enables them to compare candidates for the various explanations.

D. Robustness Considerations

There are several robustness factors worth considering. Monthly aggregation indicates monthly seasonal peaks in sales and margin volatility. Although the discount sensitivity patterns are comparable across months, seasonal demand variations may either amplify or mitigate the impact on margins. Retail sales and profit distributions are right-skewed and exhibit heavy tails. The disproportionate impact of high-value transactions on aggregate outcomes can be established. Log-scale visualizations and distributional diagnostics help reduce misinterpretation. The big states, such as California, make a disproportionate contribution to total profit. The sensitivity analyses without dominant states show that the effect of the discount threshold persists even after eliminating the high-volume states. Micro-level heterogeneity is evident in high loss rates within some subcategories. Aggregated margins should then be interpreted alongside transaction-level dispersion measures. The considerations help strengthen the belief that the main results, especially the nonlinear discount-profit relationship, are structurally sound rather than the product of aggregation.

E. Limitations

There are several things to be expressly noted. Firstly, the Sample Superstore dataset is purely illustrative and is used for illustration and reproducibility. The patterns described in this study are empirical and cannot be expected to be common across all retail stores. Secondly, the study does not make any causal statements. Visual patterns depicted by the framework are not meant to represent causal mechanisms. Thirdly, the study does not present a user study, so statements of interpretability are not based on user performance but rather on design logic and demonstrating the benchmark. Fourth, the framework has not yet been tested on several datasets, and the extent to which it can be applied across contexts remains undetermined. Fifth, predictive extensions are not within the current empirical scope and have therefore been excluded from the main analysis. Sixth, as with the other elements of the framework, there is an interpretive risk of aggregation bias, since it is based on grouped summaries at various levels of hierarchy. Lastly, the loss pockets criterion employed is not formally optimized and is descriptive and benchmark-oriented.

VI. Conclusion and Future Work

Several restrictions should be emphasized. The Sample Superstore dataset is a demonstrative dataset for illustration and reproducibility purposes only. The empirical patterns described here apply only to the dataset and cannot be assumed to be representative of retail in general. Second, this study does not make causal statements. The visual patterns displayed in the framework are not necessarily causal processes, but rather co-occurrence and analytical salience. Third, there is no user study included in the present study; thus, the interpretive claims are not based on user performance but rather on design logic and the demonstration of the benchmark. Fourth, the framework has not yet been tested on more than one dataset, and its generalizability across domains has yet to be determined. Fifth, predictive extensions are beyond the scope of the current empirical analysis and are not included in the main analysis. Sixth, because grouped summaries are used at multiple hierarchical levels, aggregation bias must be considered when interpreting the results. Lastly, the materiality criterion for loss pockets is not formally optimized and descriptive/benchmark based.

This study can be advanced in future research in several ways. Gradient boosting and causal modeling, as predictive methods, may also be added to equip the framework to recognize loss-risk trends and facilitate more sophisticated decision analysis. In general, this study demonstrates that the proposed integrated visual analytics framework is a structured method for the diagnostic analysis of retail performance, price review, and portfolio, as well as a basis for further evolution of retail decision support systems. This study presents a decision-based visual analytics model using a publicly available benchmark Superstore dataset. The findings demonstrate that coordinated descriptive and diagnostic analytics, along with organized visualizations, can help uncover data-specific profitability trends, loss zones, and risk structures based on discounts. The findings are to be understood in terms of the illustrative dataset used in the study. The framework should be tested in the future on a variety of datasets and retail environments, and expanded through predictive modeling and user feedback on the use of the decision support.

Data Availability Statement

The dataset is publicly accessible and was used in this study in accordance with its publicly available distribution for research and demonstration purposes, provided by V. Chowdhury, available online: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final (Accessed on 25 January 2026).

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Z. Atan, D. Honhon, and X. A. Pan, “Displaying and discounting perishables: Impact on retail profits and waste,” Management Science, 2025, 1–18. https://doi.org/10.1287/mnsc.2023.00316.

[2] S. Hjelle, P. Mikalef, N. Altwaijry, and V. Parida, “Organizational decision making and analytics: An experimental study on dashboard visualizations,” Information & Management, 2024, 61, 104011. https://doi.org/10.1016/j.im.2024.104011

[3] C. Burnay, M. Lega, and S. Bouraga, “Business intelligence and cognitive loads: Proposition of a dashboard adoption model,” Data & Knowledge Engineering, 2024, 152, 102310. https://doi.org/10.1016/j.datak.2024.102310

[4] L. Riesenegger and A. Hübner, “Balancing profitability and waste reduction: optimizing markdown policies for retail inventory,” International Journal of Production Research, 2026, 1-26. https://doi.org/10.1080/00207543.2025.2603570

[5] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, et al., “State of the art of visual analytics for explainable deep learning,” in Computer Graphics Forum, 2023, 319-355. https://doi.org/10.1111/cgf.14733

[6] R. C. Basole, H. Park, and C. D. Seuss, “Complex business ecosystem intelligence using AI-powered visual analytics,” Decision Support Systems, 2024, 178, 114133. https://doi.org/10.1016/j.dss.2023.114133

[7] M. R. Islam, S. Akter, L. Islam, I. Razzak, X. Wang, and G. Xu, "Strategies for evaluating visual analytics systems: A systematic review and new perspectives," Information Visualization, 2024, 23, 84-101. https://doi.org/10.1177/14738716231212568

[8] V. Chowdhury, "Superstore Dataset," Kaggle, 2021. [Online] Available at: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final (accessed on 15 February 2026)

[9] O. M. Yigitbasioglu and O. Velcu, "A review of dashboards in performance management: Implications for design and research," International Journal of Accounting Information Systems, 2012, 13, 41-59. https://doi.org/10.1016/j.accinf.2011.08.002

[10] C. Burnay, S. Bouraga, and M. Lega, "When Dashboard's Content Becomes a Barrier-Exploring the Effects of Cognitive Overloads on BI Adoption," in International Conference on Research Challenges in Information Science, 2023, 435-451. https://doi.org/10.1007/978-3-031-33080-3_26

[11] C. Schmidt, "MD Anderson breaks with IBM Watson, raising questions about artificial intelligence in oncology", Journal of the National Cancer Institute -Oxford University Press, 2017, 109(5), djx113. https://doi.org/10.1093/jnci/djx113

[12] P. Mikalef, V. A. Framnes, F. Danielsen, J. Krogstie, and D. Olsen, "Big data analytics capability: antecedents and business value," 2017. PACIS 2017 Proceedings. 136. https://aisel.aisnet.org/pacis2017/136

[13] R. C. Basole, M. G. Russell, J. Huhtamäki, N. Rubens, K. Still, and H. Park, "Understanding business ecosystem dynamics: A data-driven approach," ACM Transactions on Management Information Systems (TMIS), 2015, 6, pp. 1-32. https://doi.org/10.1145/2724730

[14] D. Chakraborty and N. P. Rana, "Impulse buying and stickiness using flow theory: a longitudinal study," Journal of Computer Information Systems, 2025, 1-15. https://doi.org/10.1080/08874417.2025.2547970

[15] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," IEEE Transactions on Visualization and Computer Graphics, 2018, 25, 2674-2693. https://doi.org/10.1109/TVCG.2018.2843369

[16] F. Doshi-Velez and B. Kim, "Considerations for evaluation and generalization in interpretable machine learning," in Explainable and interpretable models in computer vision and machine learning, Springer Cham, 2018, 3-17. https://doi.org/10.1007/978-3-319-98131-4_1

[17] Q. Li, P. Yu, and X. Wu, "Shelf life extending packaging, inventory control and grocery retailing," Production and Operations Management, 2017, 26.1369-1382. https://doi.org/10.1111/poms.12692

[18] J. Lim, I. S. Currim, and R. L. Andrews, "Consumer heterogeneity in the longer-term effects of price promotions," International Journal of Research in Marketing, 2005, 22, 441-457. https://doi.org/10.1016/j.ijresmar.2005.09.006

[19] T. H. Bijmolt, H. J. Van Heerde, and R. G. Pieters, "New empirical generalizations on the determinants of price elasticity," Journal of Marketing Research, 2005, 42, 141-156. https://doi.org/10.1509/jmkr.42.2.141.62296

[20] A. H. H. K. J. Wollenburg, "Last mile fulfillment and distribution in omni-channel grocery retailing: A strategic planning framework." International Journal of Retail & Distribution Management, 2016, 44(3), 2016. https://doi.org/10.1108/IJRDM-11-2014-0154

[21] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," Applied Soft Computing, 2011, 2664-2675. https://doi.org/10.1016/j.asoc.2010.10.015

[22] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," International Journal of Forecasting, 2021, 37, 388-427. https://doi.org/10.1016/j.ijforecast.2020.06.008

[23] E. Brynjolfsson and K. McElheran, "Data in action: data-driven decision making and predictive analytics in US manufacturing," Rotman School of Management Working Paper, 2019. http://dx.doi.org/10.2139/ssrn.3422397

[24] C. Ware, Information visualization: perception for design: Morgan Kaufmann, 2019. Morgan Kaufmann Publishers, Elsevier.Sansome Street, Suite 400, San Francisco, CA 94111. https://bravian.edu.pk/Images/Orignal/book.pdf

[25] D. Keim and L. Zhang, "Solving problems with visual analytics: Challenges and applications," in Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, 2011, 1-4. https://doi.org/10.1145/2024288.2024290

[26] W. Cui, "Visual analytics: A comprehensive overview," IEEE Access, 2019, 7, 81555-81573. https://doi.org/10.1109/ACCESS.2019.2923736

[27] S. Allam, "Exploratory Study for Big Data Visualization in the Internet of Things," Sudhir Allam, Exploratory study for big data visualization in the internet of things", International Journal of Creative Research Thoughts (IJCRT), ISSN, 2017, 2320-2882.

[28] C. M. Olszak, "Business intelligence systems for innovative development of organizations," Procedia Computer Science, 2022, 207, 1754-1762. https://doi.org/10.1016/j.procs.2022.09.233

Fostering Programming Interest Through Block Coding in Young Learners

Dorjpalam Tserendejid[1]*, Khajidmaa Otgonbaatar[2], Magsarjav Bataa[3]
Department for Education Policy, Mongolian National Institute for Educational Research, Ulaanbaatar, Mongolia[1]
Academic Secretary, Mongolian National Institute for Educational Research, Ulaanbaatar, Mongolia[2]
Department of Information and Computer Sciences, National University of Mongolia, Ulaanbaatar, Mongolia[3]

***Abstract*—Coding has become a fundamental skill for cultivating computational thinking in young learners. This study aims to examine the impact of coding education on programming interest among fifth-grade students in Mongolia with no prior coding experience. A 16-week pilot course using block-based coding activities on Code.org involved 195 5th-grade students across three schools. A pre-test/post-test design compared an experimental group ($N_{experimental}$=95) with a control group ($N_{control}$=100). Results indicate significant improvements in coding skills (Cohen's d = 0.84) and computational thinking, with average programming skill ratings increasing from 5.19 to 7.07 out of 10. Qualitative data show heightened engagement, perceived creativity, and motivation, with students describing coding as gamified activities alongside an increase in self-rated programming interest. Moreover, post-intervention interviews reveal emerging awareness of gender inclusivity in programming. These findings suggest that structured block-based coding can effectively enhance computational thinking and interest in programming among young learners, providing evidence to support early integration of programming in primary education curricula.**

***Keywords*—Computational thinking; problem-solving; block-based coding; coding education; programming interest**

I. Introduction

Today, coding has become a fundamental skill driven by technological advancements across all aspects of learning. The concept of computational thinking can be viewed as a compilation of problem-solving competencies that are intrinsically linked to the tenets of computer science [1]. There are both economic and developmental motivations for commencing computer science education in early childhood. Research demonstrates that early educational interventions in programming lead to reduced costs and more sustainable outcomes than those initiated later in a child's academic journey [2], [3]. With the emergence of the sixth information technology revolution, defined by artificial intelligence, the ability to code has become a vital competitive edge for today's workforce. Young students are enthusiastic about quickly learning basic digital skills to adapt to the evolving information society [4]. Additionally, [5] highlighted that computational thinking is an essential skill set that everyone should be motivated to acquire and apply in their everyday lives. Moreover, [6] proposed that Computational Thinking (CT) has existed since the 1950s, originally termed 'algorithmic thinking', which involves employing a systematic and precise sequence of steps to address a problem, utilizing a computer when suitable for this purpose.

Through the digital landscape, computational thinking (CT) and coding have become critical foundational skills, now recognized as essential parts of national curricula in many education systems. CT refers to mental processes such as breaking down problems, recognizing patterns, designing algorithms, and abstracting solutions in ways that can be automated using technology [5]. Coding serves as a practical application of these skills, allowing learners to express ideas through structured, programmable instructions, promoting problem-solving, logical reasoning, and digital fluency [7], [8].

Although there is no consensus among academics on a definitive interpretation of computational thinking (CT), Wing describes it as the process of problem-solving, system design, and the concept of human behavior, drawing from the essential principles of Computer Science. She notes that computational thinking is not merely programming but rather involves a broader scope that includes thinking at multiple levels of abstraction [5]. In addition, [5] describes computational thinking as a form of analytical thought that seeks to understand human behavior, offering a structure on essential concepts such as generating system designs, addressing problems, and processing information. On one hand, studies have shown that instruction in computational thinking (CT) fosters a deeper understanding that programming is fundamentally about addressing problems, rather than simply coding, and it can bolster female students' perceptions and self-assurance in programming [9]. It is noteworthy that CT serves as an early indicator and predictor of academic success, as the CT scores have shown a strong correlation with overall academic performance [10].

Several countries, including Finland, the United Kingdom, South Korea, and the United States, have adopted CT and coding in primary and secondary education as part of broader digital literacy and STEM agendas [8], [11]. Research consistently shows that introducing coding at an early age supports cognitive development, including executive function, critical thinking, and conceptual reasoning [11], [12]. Beyond cognitive gains, coding activities, especially when delivered through visual programming platforms, mobile apps, or educational robotics, enhance learner motivation, engagement, collaboration, and creativity [11], [13]. Importantly, CT is not limited to computer science alone. It supports learning across diverse subjects, including mathematics, science, language arts, and the arts, and promotes interdisciplinary problem-solving [14], [15]. These cross-cutting benefits have led to a global

*Corresponding author

shift toward embedding coding and CT within general education rather than treating them as niche or elective skills.

The Mongolian education context, [16], [17] explicitly calls for enhanced digital competencies, improved access to STEM education, and inclusive, future-oriented learning pathways. Integrating coding and CT into primary education supports these goals by preparing students with relevant skills for digital society and addressing urban–rural disparities in access to quality education. This study responds to the need for localized evidence by examining the effects of structured coding activities on the development of computational thinking in fifth-grade students in Mongolia. Through a pre-test/post-test design, the study analyzes both the statistical significance and practical impact of coding instruction, contributing to curriculum reform and global dialogue on 21st-century education.

II. Literature Review

A. Importance of Coding Education

Literature defines coding as the process by which computers transform human-understandable operations into executable algorithms or sequences of commands [18]. It has been proposed that coding can encourage self-directed learning. Additionally, it supports literacy and mathematical communication, as well as metacognitive abilities, especially when children are introduced to coding at an early age [19]. Furthermore, learning to code equips students with a novel symbolic language that enables them to develop programs and projects, thereby establishing a positive link between coding skills and creative thinking [20].

Introducing children to coding, an essential skill for digital literacy and the language of today's world, should occur in early developmental stages of children [21]. The practice of coding also enables analysis, problem-solving, and concept development by formulating algorithms that break down problems and representing these algorithms in a programming language. In this way, coding serves as a medium for children to express their ideas and perspectives [22]. Programming also offers children opportunities to “interact" with computers and allows them to learn new symbolic languages [7], [20]. Additionally, 21st-century skills are essential for individuals, with coding being considered a fundamental skill. Coding should be included among the competencies that every student ought to acquire during their educational journey [23].

On the other hand, the concept of 'coding without computers' relates to educational experiences where students engage in lessons known as 'unplugged'. In these lessons, the emphasis is not on algorithmic reasoning, computer interactions, or database management. Instead, students often use pencils and paper or physical manipulatives. These thoughtfully crafted kinesthetic activities allow students to understand intricate concepts through cards, games, and puzzles. Such experiences are pertinent to their lives and promote the learning of fundamental information technology principles [24]. In addition, numerous studies noted that graphical environments can support cognitive growth in children and are more effective for teaching programming to younger age groups [25], [26].

B. Unpacking Interest in Creative Programming

In terms of creative learning, constructivism is recognized as a significant concept in today's educational sphere. Yet only a small fraction of classrooms genuinely encourages students to approach their schoolwork as a creative act of knowledge construction. The goal of these initiatives is to create a classroom environment that functions as a community for building knowledge [27]. As the integrated mode is recognized as a more imaginative processing style, the authors advised that computer programming instructors should prioritize software design, which utilizes an integrated mode [28]. Moreover, a variety of computer programs are accessible. To enhance the cultivation of creative thought, these programs should function as tools for cognitive processes. We must exercise caution against providing a haphazard assortment of programs or tools that lack internal coherence [29].

Engaging in creative coding requires diverse problem-solving competencies that leverage computer algorithms, often referred to as “computational thinking” [5]. On the one hand, Computational thinking (CT) refers to a method of problem-solving that encompasses the formulation of problems; the logical organization and analysis of data; the representation of data via abstractions; the identification and automation of solutions through algorithmic reasoning; the analysis and implementation of potential solutions; and the generalization and transfer of the problem-solving process [30], [31].

III. Methodology

A mixed-methods parallel design was employed, integrating quantitative and qualitative approaches with equal weighting. The structural workflow of this mixed-methods parallel design is visually illustrated in Fig. 1.

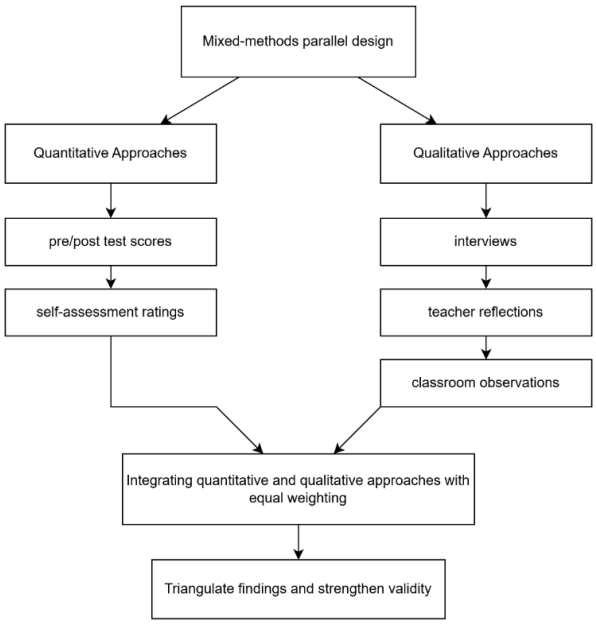


Fig. 1. Mixed-methods parallel design of the study.

This design allowed the combination of statistical analysis and thematic interpretation to triangulate findings and strengthen validity [32]. Quantitative data included pre/post

test scores and self-assessment ratings, while qualitative data were derived from interviews, teacher reflections, and classroom observations. Qualitative data from interviews, classroom observations, and teacher reflections were collected concurrently to enrich interpretation and triangulate findings. For the qualitative phase, we employed a purposive stratified sampling method to select 15 participants from the experimental group ($n_{experimental}$=95). This method was chosen to ensure a diverse representation of gender, geographic location (urban vs. rural), and varying levels of coding progression observed during the 16-week intervention. Existing classroom groups were used without random assignment due to practical school-level constraints.

Both methods were implemented concurrently and analyzed independently before being merged in a comprehensive interpretation. This approach corroborates quantitative data with rich descriptive insights from classroom practices and interviews, etc.

A. Research Context and Sampling

This study was conducted as part of a national pilot project designed to integrate computer science education into the Mongolian primary curriculum. The research focused on fifth-grade students who had no prior coding experience. The objective was to examine how structured, block-based coding activities influence computational thinking and interest in programming. The research data refers to 195 students in 5th grade from three public schools, consisting of two rural schools and one urban school, including 97 males and 98 females, who possessed no previous coding knowledge or skills and were being introduced to the subject for the first time. Teachers enrolled in the pilot lessons had limited coding experience and pedagogical confidence. However, they received pre-study training and continuous pedagogical support. Ethical approval and informed consent were obtained from all participants and their guardians before data collection.

B. Teacher Profile and Preparation

Fifteen teachers participated in the pilot implementation (see Table I). The majority were female (86.7%) and held a bachelor's degree (66.7%). Before classroom implementation, all teachers attended an 8-hour professional development workshop focused on Code.org pedagogy, computational thinking concepts, and strategies for facilitating creative, student-centered learning. Teachers were provided with lesson plans, teaching aids, and access to a monitoring platform to track student progress and ensure consistent delivery across schools.

C. Instruments and Outcome Measures

To evaluate the intervention, two distinct quantitative assessment tools were deployed at baseline (pre-test) and post-intervention (post-test).

1) Computational thinking performance assessment: This 10-point instrument was adapted from the verified Code.org Course C diagnostic metrics and validated primary school block-coding tasks established by relevant literature. It comprises 10 items (scored 0 or 1 point each, yielding a 0–10 composite scale) that assess five core computational thinking sub-constructs: algorithmic design (2 items), pattern recognition (2 items), loops/repetition logic (2 items), abstraction (2 items), and debugging logic (2 items). A pilot testing phase involving an independent cohort of fifth-grade students (N = 25) yielded a Cronbach's alpha of $\alpha = 0.81$, indicating robust internal consistency. Construct validity was confirmed through evaluation by an expert panel comprising two educational technology professors and one senior computer science teacher.

2) Self-rated programming interest and ability scales: Student perceptions were collected through two standalone 10-point Likert-type single items ("Rate your interest in learning computer programming" and "Rate your current ability to write code/solve puzzles", where 1 = Very Low and 10 = Very High).

TABLE I. THE DEMOGRAPHIC CHARACTERISTICS OF TEACHERS

Variables	Values	Frequency	Percentage (%)
Gender	Female	13	86.7%
	Male	2	13.3%
Educational attainment	Bachelor's degree	10	66.7%
	Master's degree	5	33.3%
Age bracket	25–35	8	53.3%
	36–45	3	20.0%
	46–55	4	26.7%
Professional commitment	2–5 years	3	20.0%
	6–10 years	6	40.0%
	11–20 years	2	13.3%
	21–30 years	2	13.3%
	31–40 years	2	13.3%
Experience in teaching coding	Yes	0	0.0%
	Somewhat	5	33.3%
	No	10	66.7%

D. Qualitative Data Collection and Analysis Rigor

The qualitative strand utilized semi-structured interview protocols alongside structured classroom observations. Interviews were conducted in the native Mongolian language by the primary research team to eliminate communication barriers. All recorded audio files were transcribed and translated into English using a back-translation protocol. The text was analyzed using inductive thematic analysis. Two researchers independently coded the transcripts to identify repeated semantic patterns regarding student enjoyment, confidence, and equity. Inter-rater reliability was verified using Cohen's Kappa *(k)*, yielding an initial agreement of $k = 0.83$. Minor coding discrepancies were resolved through iterative team consensus meetings. Final frequencies (e.g., 11 out of 15 participants noting high enjoyment) represent systematic code distribution across the entire sample rather than illustrative anecdotes. Table II presents the demographic data and backgrounds of participating students.

TABLE II. DEMOGRAPHIC DATA AND BACKGROUNDS OF PARTICIPATING STUDENTS.

Demographic and Personal Information	Values	Frequency	Percentage (%)
Gender of students	Male	97	49.7%
	Female	98	50.3%
Student age	10 years old	189	96.9%
	11 years old and above	6	3.1%
Previous coding experience	Yes	5	2.6%
	No	190	97.4%
Parents work in IT-related jobs	Yes	5	2.6%
	No	190	97.4%
Awareness of platforms	Khan Academy	3	1.5%
	Scratch	1	0.5%
	Code.org	2	1.0%

E. Course Design and Delivery

The 16-week pilot course comprised 16 lessons, each lasting approximately 40 minutes. The structure of each lesson followed a consistent pattern:

1) Introduction (10 minutes): Presentation of the lesson goal and demonstration of a simple concept.

2) Guided coding practice (20 minutes): Students completed block-based coding challenges on Code.org using computers or tablets and extra coding tasks.

3) Reflection and discussion (10 minutes): Students discussed solutions, errors, and strategies to reinforce computational thinking and creativity.

The intervention combined computer-based coding and unplugged activities, such as card sequencing, paper-based tasks, and movement-based problem-solving, to reinforce algorithmic reasoning without reliance on digital tools. This blended approach allowed students to grasp core programming logic through both physical and digital representations. Throughout the intervention, teachers maintained reflective notes, and researchers conducted classroom observations to record behavioral engagement and problem-solving interactions.

The 16-week educational intervention was designed to isolate, emphasize, and evaluate block-based programming skills by mapping the digital environments of Code.org Course C and Course D (2024 editions) directly into the core coding chapters of the official Mongolian 5th-grade Computer Science textbook (used only in curriculum pilot). While the textbook introduces physical computational hardware and foundational robotics elements in its introductory units, this pilot study intentionally focused exclusively on the algorithmic logic and programming modules to systematically evaluate progress in students' computational thinking over the course.

The curriculum structure was split into four interconnected 4-week thematic blocks. To ensure a blended learning approach, every theoretical programming paradigm introduced in the national textbook was immediately applied and tested through matching interactive, gamified puzzles on Code.org. The specific chronological layout, matching local textbook units, practical student block operations, and targeted computational thinking dimensions are comprehensively outlined in Table III.

TABLE III. SYSTEMATIZED 16-WEEK EXPERIMENTAL PROGRAMMING TIMELINE MAPPED TO MONGOLIAN 5TH GRADE TEXTBOOK MODULES

Phase / Timeline	Textbook Unit & Chapter Alignment	Corresponding Code.org Modules	Practical Student Programming Tasks & Core Operations	Targeted Computational Thinking (CT) Competency
Weeks 1–4	(Unit 2: Characters and Movement - Lessons 2.1, 2.2)	Course C: Lessons 1–4: Digital Citizenship & Sequencing in Maze	Students constructed explicit sequential command chains to move sprites autonomously. Handled event-driven logic by mapping interactions to physical action blocks (e.g., programming character speech outputs on clicking events).	Algorithmic Thinking & Decomposition: Breaking single overarching motion routes into individual directional vectors.
Weeks 5–8	(Unit 2: Optimization - Lesson: Sequencing & Loops)	Course C: Lessons 5–8: Loops in Laurel & Artist	Students identified geometric repetitions in design files. Programmed nested "Repeat" parameters to construct complex grid arrays and multi-sided shapes without writing redundant lines of individual code.	Pattern Recognition & Code Efficiency: Utilizing structural patterns to minimize command redundancy via loops.
Weeks 9–12	(Unit 3: Variables & Game Design - Lessons 3.1–3.3)	Course D: Lessons 9–11: Conditionals & Scores in Bee/Farmer	Students initialized dynamic parameters to track game actions (e.g., / score variables). Built condition checks ("If/Else") forcing characters to gather assets only when strict environmental variables were met.	Conditional Logic & Data Tracking: Designing rule-based branching paths and tracking fluid states via data variables.
Weeks 13–16	(Unit 4: Functions & Optimization - Lessons 4.1, 4.2)	Course D: Lessons 12–14: Functions in Minecraft & End Project	Students audited broken scripts to isolate hidden logical failures (Debugging). Wrapped complex, recurring code instructions into distinct named macro modules to assemble an original capstone project.	Abstraction and Debugging: Encapsulating complex sub-routines into clean functions; diagnosing system failures methodically.

IV. RESULTS

A. Coding Experiment Outcomes

The box plot (Fig. 2) illustrates the changes in score distribution after the intervention, emphasizing trends that extend beyond mere mean differences. In the experimental group, the post-test scores show a higher median and a reduced interquartile range, indicating enhanced consistency in learning outcomes. This trend is consistent with previous studies indicating that coding activities can improve not only individual computational thinking (CT) skills but also foster more equitable performance among students [33], [15].

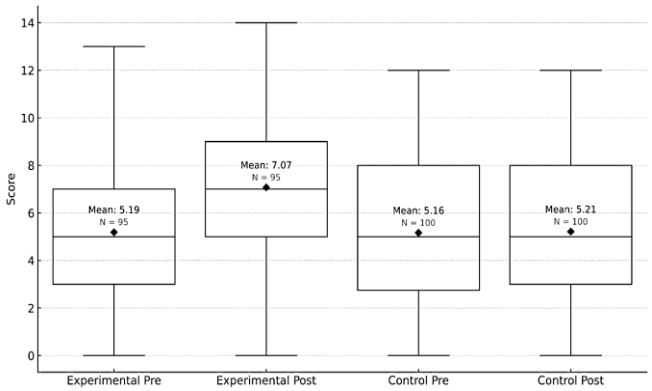


Fig. 2. Box plot of pre- and post-intervention scores in experimental and control groups.

In contrast, the control group demonstrates only a slight upward shift in the mean, with overlapping interquartile ranges both before and after the test. This suggests that general instruction alone may yield minimal or uneven learning gains in CT, consistent with studies indicating that explicit, structured computational tasks are more effective than implicit exposure [34], [35]. The absence of extreme outliers in both groups further supports the reliability of the intervention's effect. Significantly, the experimental group's decrease in variability suggests that the coding program was more effective in assisting lower-performing students - a pattern linked to well-structured CT interventions [36].

These findings support the interpretation that coding activities, when systematically implemented, lead to both higher and more consistent CT performance among primary school learners.

B. Coding Ability Progression (Hypothesis 1)

To evaluate the progression of students' coding capabilities following the 16-week intervention, quantitative tracking metrics were analyzed. This section details the statistical shifts in objective computing performance alongside subjective self-efficacy developments across both cohorts, as highlighted in the metrics below. The detailed post-intervention evaluation metrics and assessment score breakdowns are summarized in Fig. 3.

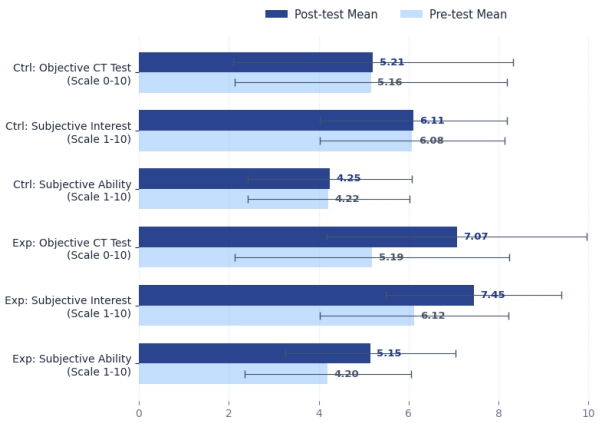


Fig. 3. Assessment score metrics

C. Coding Experiment Outcomes

Before the intervention, baseline equivalence between the groups was statistically verified. A two-sample independent t-test conducted on the pre-test scores showed a statistically insignificant difference between the experimental group (M = 5.19, SD = 2.98) and the control group (M = 5.16, SD = 3.02; t (195) = 0.07, p = 0.944), validating their baseline comparability.

Following the 16-week intervention, a paired sample t-test was conducted to examine intra-group progress from pre-test to post-test. Because zero student attritions occurred during the course, the degrees of freedom (*df*) correspond directly to n – 1 for each group. For the experimental group (n = 95), computational performance scores increased significantly from M = 5.19 (SD = 2.98) to M = 7.07 (SD = 3.14), yielding t (94) = -8.13, $p < 0.001$. The standardized effect size was computed using the paired-samples standardized mean difference formula:

$$d = \frac{M_{post} - M_{pre}}{SD_{difference}}$$

where, M_{post} and M_{pre} represent the post-test and pre-test means, respectively, and $SD_{difference}$ represents the standard deviation of the difference scores.

These findings indicate that structured block-based coding activities significantly improved students' computational thinking and coding performance. To verify the specific areas of student improvement behind these mean differences, an item-level analysis was conducted. As explicitly mapped across the curriculum topics in Fig. 4, the experimental pilot group showed a substantial accuracy advantage over the control class group, particularly in advanced block logic tasks.

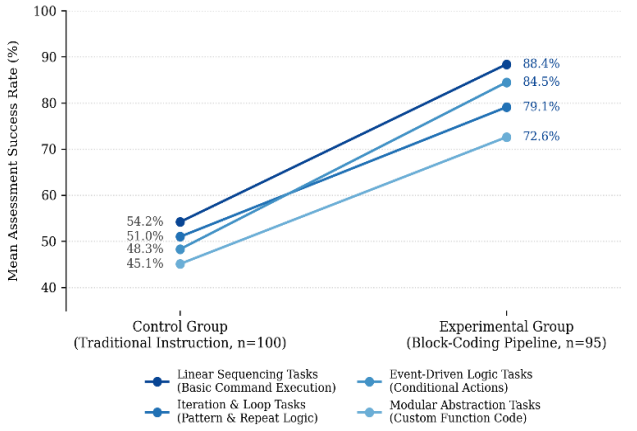


Fig. 4. Explicit performance accuracy tracking across core coding tasks, comparing the mean success rates of the Experimental Pilot Group (n = 95) and the Traditional Class Group (n = 100).

In contrast, the control group ($n_{control}$ = 100) exhibited a statistically negligible change from pre-test (M = 5.16, SD = 3.02) to post-test (M = 5.21, SD = 3.11), with t (99) = 0.74, p = 0.456 and a negligible effect size (Cohen's *d* = 0.07).

D. Programming Interest Gain Scores (Post-Test – Pre-Test)

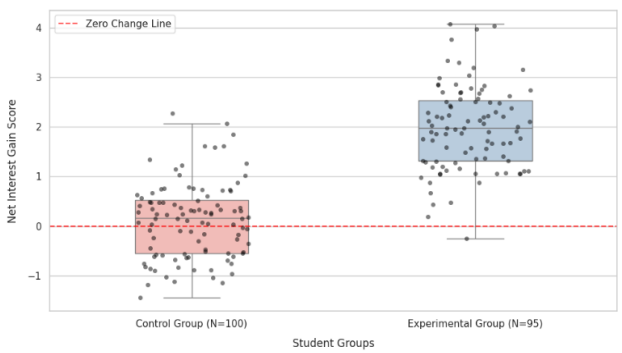


Fig. 5. Programming interest gain scores between the experimental and control groups.

To see exactly how the program affected each student, we calculated personal progress by subtracting pre-test scores from post-test scores. As shown in Fig. 5, the results are completely different. In the Control group, almost everyone's score is crowded right around the zero line, meaning regular classes did not really change their interest in coding.

In contrast, almost every student in the Experimental group shifted well above the zero line. Even the middle half of the experimental group sits completely above the control group's highest concentration. This visually confirms our strong overall effect size (Cohen's d = 0.84) and proves that the coding activities systematically increased programming interest across the entire classroom rather than shifting it via a few standout students.

E. Programming Interest and Motivation (Hypothesis 2)

Changes in student self-rated programming interest are described in the metrics below and mapped visually in Fig. 6.

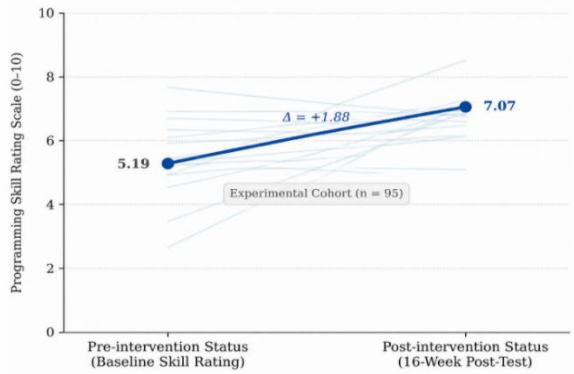


Fig. 6. Longitudinal trajectory shift mapping the baseline programming self-efficacy ratings against post-intervention evaluation metrics within the experimental cohort (n = 95).

TABLE IV. QUALITATIVE EVIDENCE OF INCREASED INTEREST IN PROGRAMMING.

Theme	Frequency	Example Student Quotes
Enjoyment of coding	11/15	"Coding is like playing a game."
Perceived creativity	9/15	"I can make easier stories and games."
Increased confidence	8/15	"I want to learn more coding after this class."

Quantitative and qualitative data jointly support Hypothesis 2. Students not only improved in self-rated programming ability (+1.88) but also developed positive attitudes toward coding, describing it as fun, creative, and motivating. The data show that students who engaged in Code.org block-based coding showed significant improvements in coding scores and increased interest in programming, as observed in interviews and classroom behavior. Qualitative responses revealed that students described coding as "fun", "creative", and "like playing games", indicating increased motivation and engagement. Together, the significant improvement in test scores and self-reported enthusiasm support Hypothesis 2. Table IV shows qualitative evidence of increased interest in programming.

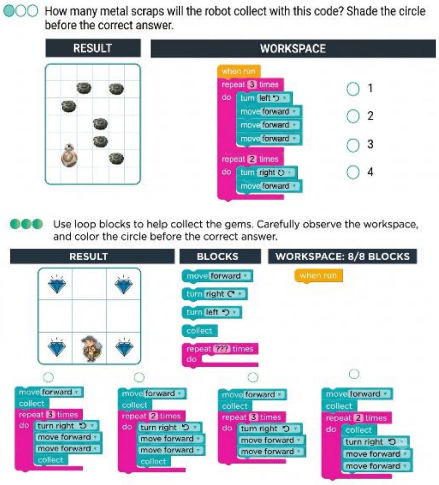


Fig. 7. Supplementary reference items from the post-intervention assessment instrument. (A) Linear grid sequencing task (Q2), (B) Iterative loop optimization task (Q9), and (C) Asynchronous event-driven matching task (Q7).

Together, the significant improvement in test scores and self-reported enthusiasm support Hypothesis 2. To visually confirm the nature of this data improvement, Fig. 7 provides supplementary reference samples of the task logic successfully parsed by the pilot cohort. The observed score improvements reflect students' growing ability to use more advanced computational thinking concepts, including loops and event-based programming structures.

F. Gender Awareness and Inclusivity (Hypothesis 3)

Although not universal, interview data indicate a shift toward gender-equitable views, providing partial support for Hypothesis 3. Female students showed increased confidence and interest, reflecting the potential of early coding instruction. Furthermore, post-intervention interviews revealed that several students recognized gender-related themes, with female students acknowledging increased confidence in coding skills and noting that "both girls and boys can be good at programming". Although initial surveys reflected limited awareness of gender-related stereotypes in computer science, the course prompted discussions about inclusivity. Teachers also reported observing a reduction in gendered perceptions of coding as an activity "more suited for boys". These findings provide partial support for Hypothesis 3: students' improved grasp of Computer Science concepts includes an emerging

awareness of gender equity and challenges to existing stereotypes, although further targeted instruction may be needed to reinforce these outcomes. Table V presents the gender-related awareness indicators (post-intervention interviews).

TABLE V. GENDER-RELATED AWARENESS INDICATORS (POST-INTERVENTION INTERVIEWS).

Indicator	Frequency (n=15 interviews)	Example Quotes
Acknowledgment that girls can code	6	"Girls can be good coders like boys."
Expression of inclusivity in coding	4	"Everyone should try coding; it's for everyone."
Recognition of existing stereotypes	2	"I thought coding was just for boys before."

V. DISCUSSION

The study demonstrates that structured, block-based coding education significantly enhances computational thinking, motivation, and further interest in programming among fifth-grade students with no prior coding experience. Quantitative results showed a large performance improvement (Cohen's d = 0.84), confirming the intervention's effectiveness and aligning with findings from [12], [13].

Qualitative responses further reveal that coding is perceived as *"fun", "creative",* and *"like playing games"*. This indicates that interactive, game-like design elements in block-based coding environments reduce cognitive barriers and foster intrinsic motivation. Such experiences transform programming from an abstract, syntax-heavy subject into an accessible and enjoyable learning process. Post-intervention discussions also revealed growing gender inclusivity, as both male and female students expressed that *"anyone can code."* Although awareness of stereotypes was not universal, the trend toward inclusivity suggests that early coding education can play a meaningful role in addressing gender disparities in STEM participation [11].

Within the context of [16], [17], these findings provide actionable evidence for integrating computational thinking and coding into the primary education curriculum. The visual, low-barrier coding tools such as Code.org can reduce urban–rural learning gaps and foster equitable digital participation.

The qualitative interview findings provide preliminary insights into student engagement and perceptions. Regarding gendered views of computing, 6 out of the 15 interviewed participants (40%) explicitly verbalized that girls and boys possess equal capability in programming tasks. While these responses indicate positive individual experiences among female students and point to potential self-confidence gains within the sample, this exploratory finding cannot be extrapolated as a sweeping attitudinal shift across the wider student population. Because baseline gender-role stereotype measures were not quantitatively tracked from the outset, these expressions are interpreted purely as emerging individual awareness among the interviewees. This highlights a critical area that requires dedicated, structured tracking in future longitudinal research.

Therefore, future research should build on this pilot by including larger, more diverse samples, longitudinal data, and teacher training components. More investigation into socio-cultural and contextual factors - such as parental influence, language barriers, and teacher digital competence - will enhance the scalability and sustainability of computer education reform.

VI. CONCLUSION

This study demonstrates that the strategic integration of block-based coding activities into primary education effectively develops foundational computational thinking, operational problem-solving, and a heightened interest in computer programming among young learners with no prior experience. Over the course of a 16-week intervention utilizing Code.org, fifth-grade students in Mongolia demonstrated statistically significant and practically substantial gains in objective coding metrics, while concurrently shifting their subjective perceptions of computer science from an abstract, syntax-heavy domain into an accessible, creative endeavor. Furthermore, the study provides preliminary evidence indicating that early exposure can foster individual awareness of gender equity in technology, helping minimize stereotypical biases before they solidify in later academic stages. Ultimately, providing structured, low-barrier visual programming pathways within primary school frameworks serves as a practical sandbox for building equitable digital competencies and offers localized empirical evidence to support ongoing nationwide computer science curriculum reforms.

However, while some promising awareness regarding gender emerged post-intervention, the results suggest that brief interventions are not a cure-all; additional targeted curricular measures are necessary to enforce long-term gender-equitable attitudes. To address the limitations of this pilot study, future research should recruit a larger, more diverse student sample spanning multiple regional backgrounds to enhance generalizability. Additionally, despite high modern exposure to artificial intelligence (AI) technologies, further longitudinal tracking is recommended to evaluate how student attitudes, computational thinking retention, and intrinsic curiosity evolve alongside national educational policies.

ACKNOWLEDGMENT

The authors wish to thank all the students, teachers, parents, and training managers who gave consent and participated in the study. The authors also thank the teachers who shared their feedback and suggestions for improving the curriculum content and inclusivity.

ETHICAL CONSIDERATIONS

The study involved minor participants (fifth-grade students). The study protocol was approved by the Research Committee of the Mongolian National Institute for Educational Research. All procedures performed were conducted in accordance with the ethical standards of the institutional research committee and the national guidelines for educational research involving minors. Written informed consent was obtained from the school principals and the students' parents or legal guardians. Additionally, verbal assent was secured before administering assessments, observations, or interviews. All

student data were completely anonymized using unique numeric identifiers, and participants were informed of their right to withdraw from the activities at any stage without academic penalty.

REFERENCES

[1] P. Curzon, J. Black, L. R. Meagher, and P. McOwan, "cs4fn. org: Enthusing students about computer science," Proceedings of Informatics Education Europe IV, pp. 73–80, 2009.

[2] F. Cunha and J. Heckman, "The technology of skill formation," *American Economic Review*, vol. 97, no. 2, pp. 31–47, 2007.

[3] J. Heckman and D. Masterov, "The productivity argument for investing in young children," *Review of Agricultural Economics*, vol. 29, no. 3, pp. 446–493, 2007.

[4] S. Popat and L. Starkey, "Learning to code or coding to learn? A systematic review," *Computers & Education*, vol. 128, pp. 365–376, 2019.

[5] J. M. Wing, "Computational thinking," Communications of the ACM, vol. 49, no. 3, pp. 33–35, Mar. 2006.

[6] P. J. Denning, "The profession of IT: The problem of computation," *Communications of the ACM*, vol. 52, no. 9, pp. 132–141, 2009.

[7] M.U. Bers, Coding as a Playground: Programming and Computational Thinking in the Early Childhood Classroom, New York, NY, USA: Routledge, 2018.

[8] M. Resnick, Lifelong Kindergarten: Cultivating Creativity through Projects, Passion, Peers, and Play. Cambridge, MA, USA: MIT Press, 2017.

[9] R. Davies, "Gender and computing: Understanding female students' attitudes and self-efficacy in programming," *Computers & Education*, vol. 50, no. 2, pp. 430–440, 2008.

[10] H. Haddad and F. Kalaani, "The predictive role of computational thinking skills on academic performance," *Education and Information Technologies*, vol. 20, no. 4, pp. 567–578, 2015.

[11] S. Papadakis, "Apps to promote computational thinking and coding skills to young-age children: A pedagogical challenge for the 21st century learners," *Educational Process: International Journal*, vol. 11, no. 1, pp. 7–13, 2022. doi: 10.22521/edupij.2022.111.1.

[12] Ş. Metin, M. Başaran, and D. Kalyenci, "Examining coding skills of five-year-old children," *Pedagogical Research*, vol. 8, no. 2, Article em0154, 2023.

[13] A. Strawhacker and M. U. Bers, "What they learn when they learn coding: Investigating cognitive domains and computer programming knowledge in young children," *Educational Technology Research & Development*, vol. 67, no. 3, pp. 541–575, Jun. 2019.

[14] C. Tofel-Grehl, K. A. Searle, and B. L. MacDonald, "Lighting up history: Integrating mathematics and computational thinking in the science classroom," *Science Scope*, vol. 44, no. 5, pp. 64-71, 2021.

[15] K. A. Mills, J. Cope, L. Scholes, and L. Rowe, "Coding and computational thinking across the curriculum: A review of educational outcomes," *Review of Educational Research*, vol. 94, 2024. DOI: 10.3102/00346543241241327.

[16] Government of Mongolia, *Vision-2050 Long-Term Development Policy of Mongolia*, Ulaanbaatar, Mongolia, 2020.

[17] Ministry of Education and Science of Mongolia, *Education Sector Medium-Term Development Plan (2021–2030)*, Ulaanbaatar, Mongolia, Dec. 2020.

[18] P. Van Roy and S. Haridi, *Concepts, Techniques, and Models of Computer Programming*, Cambridge, MA, USA: MIT Press, 2004.

[19] D. H. Clements, "Curriculum research: Toward a framework for examining early mathematics learning," Journal for Research in Mathematics Education, vol. 33, no. 2, pp. 111–136, 2002.

[20] F. Wang, Y. Li, and X. Zhang, "The impact of coding on students' creative thinking and problem-solving skills," *Computers & Education*, vol. 170, p. 104220, 2021.

[21] G. Futschek and R. Moschitz, "Algorithmic thinking: The key for understanding computer science," in *Proc. 5th Baltic Sea Conf. Comput. Educ. Res.*, Koli, Finland, 2010, pp. 159–168. doi: 10.1145/1867651.1867677.

[22] M. U. Bers, Blocks to Robots: Learning with Technology in the Early Childhood Classroom, New York, NY, USA: Teachers College Press, 2008.

[23] J. Voogt, N. Fisser, R. Good, P. Mishra, and K. Yadav, "Computational thinking in compulsory education: Towards an agenda for research and practice," *Education and Information Technologies*, vol. 20, no. 4, pp. 715–728, 2015.

[24] F. Kalelioğlu and O. Keskinkılıç, "A study on the effects of unplugged programming activities on computational thinking skills of young learners," *Journal of Educational Computing Research*, vol. 55, no. 7, pp. 909–933, 2017.

[25] F. Kalelioğlu, "The effects of visual programming environments on young learners' computational thinking skills," *Computers & Education*, vol. 87, pp. 1–12, 2015.

[26] S. Zhang and J. Nouri, "Teaching programming in primary schools: Effectiveness of graphical programming environments," *Education and Information Technologies*, vol. 24, no. 2, pp. 1231–1249, 2019.

[27] M. Scardamalia and C. Bereiter, Higher Levels of Knowledge: Teachers' Roles in a Knowledge-Building Community, New York, NY, USA: Routledge, 1992.

[28] J. Chesson, R. McBride, and T. Cartwright, "Integrated mode processing in computer programming: Implications for software design education," *Journal of Educational Computing Research*, vol. 8, no. 3, pp. 323–338, 1992.

[29] D. H. Clements and B. K. Nastasi, "Supporting creativity in computer-based learning environments," *Journal of Educational Computing Research*, vol. 8, no. 4, pp. 451–468, 1992.

[30] V. Barr, C. Harrison, and L. Conery, "Computational thinking: A digital age skill for everyone," *Learning & Leading with Technology*, vol. 38, no. 6, pp. 20–23, 2011.

[31] A. Ioannidou, J. Sampson, and M. Karagiannidis, "Computational thinking: Educational perspectives and practice," *ACM SIGCSE Bulletin*, vol. 43, no. 1, pp. 322–326, 2011.

[32] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, 3rd ed., Thousand Oaks, CA, USA: SAGE Publications, 2017.

[33] S. Grover and R. Pea, "Computational thinking in K–12: A review of the state of the field," *Educational Researcher*, vol. 42, no. 1, pp. 38–43, 2013.

[34] A. Fessakis, K. Gousiou, and P. Matsinas, "Children's computational thinking and programming — Unplugged and plugged-in activities," *Computers & Education*, vol. 69, pp. 418-427, 2013.

[35] T. Soffer and A. Cohen, "Students' engagement characteristics predict success and completion of online courses," *Journal of Computer Assisted Learning*, vol. 35, no. 3, pp. 378–389, 2019. DOI: 10.1111/jcal.12340.

[36] Á. Sáez-López, E., Aguaded, and A. Martín-Márquez, "Computational thinking interventions in K–12 classrooms: Effects on student learning and equity," Computers in Human Behavior, vol. 107, pp. 105637, 2020.

MYPO-Net: A Robust Deep Learning Approach for Multi-Yoga Pose Detection and Occlusion Handling

Rehana Danial[1], Nosheen Qamar[2]*, Nosheen Sabahat[3], Faria Nazir[4], Ali Salem Bin Sama[5], Lamia Hassan Rahamatalla[6], Osman Elwasila[7], Abdulrahman Alojail[8], Marwan Abu-Zanona[9]
Department of Software Engineering, University of Management & Technology, Lahore, Pakistan[1, 2, 4]
Department of Computer Science, Forman Christian College University, Lahore, Pakistan[3]
Department of Management Information Systems-College of Business Administration,
King Faisal University, Al-Ahsa 31982, Saudi Arabia[5, 6, 7, 8, 9]

***Abstract*—Yoga has become a well-known holistic process worldwide and has been appreciated due to its physical, psychological, and injury-preventive effects. The swift development of online fitness applications has created a growing need for automated systems that can precisely identify and analyze yoga poses. The current methods, however, are limited in terms of the limited diversity of datasets, insufficient occlusion, low performance in multi-person settings, and a lack of feedback mechanisms to offer corrective feedback. In order to overcome these shortcomings, this study introduces MYPO-Net, an artificial intelligence (AI) based deep learning model that uses the efficiency of MobileNet and the classification performance of EfficientNetB0. The model is trained and tested on the Yoga82 data, with a detailed preprocessing pipeline, such as resizing, normalization, and data augmentation, to improve resilience to real-world variations. Experimental evidence shows a classification accuracy of 97.65, which is higher than a variety of baseline architectures (VGG16: 87%, InceptionV3: 82%, ResNet50: 58) and has high computational efficiency. Confusion matrix analysis shows that there is valid detection in 16 yoga position classes. The persisting issues in real-time implementation and poor image quality settings are distinguished as future work directions. MYPO-Net is a highly scalable, affordable, and open-source platform to support digital yoga teaching, fitness apps, and rehabilitative health care.**

Keywords*—Yoga pose classification; deep learning; mobilenet; efficientnetb0; occlusion handling; transfer learning; convolutional neural networks; human pose estimation

I. Introduction

Yoga is a thousand-year-old practice that began in ancient India and combines physical poses (asanas), breathing exercises (pranayama), and meditation to promote holistic health. Yoga has evolved in the last 20 years into a global fitness experience, leaving behind the traditional spiritual practice and appealing to a wide range of practitioners of all ages, fitness, and cultural backgrounds. In 2022, the yoga market was estimated to be USD 37.5 billion globally and is expected to increase at a compound annual growth rate of 9.6% until 2030 [11]. As the trend towards the use of digital health platforms, mobile fitness apps, and remote health support systems grows exponentially, the need for intelligent and automated yoga teaching systems is at an all-time high.

The traditional approach to teaching yoga requires a physical presence of a trained teacher who can monitor the body position of a student, detect postural discrepancies, and give individual corrective feedback in real-time. This kind of practice with supervision of an expert is commonly considered the safest and most effective way of learning yoga, especially among beginners who might be uninformed about the misalignments that can be hard to detect but can lead to greater injury. Nonetheless, the widening disconnect between the supply of instructors and the rising number of practitioners in the world as a barrier to accessing expert-supervised instruction has been accelerated by geographical barriers, economic factors, and the transition to home-based practice due to the COVID-19 pandemic, making expert-supervised instruction no longer affordable to millions of practitioners [4]. This situation has spurred extensive research on computer vision and deep learning models capable of automatically identifying, recognizing, and judging yoga poses without human oversight.

Deep learning and computer vision have fueled a new generation of human pose estimation frameworks with impressive performance. Convolutional neural networks (CNNs) such as ResNet [13], VGG [12], and InceptionV3 [14] have established strong benchmarks for image-based recognition tasks, while dedicated pose estimation models, including OpenPose, MoveNet, BlazePose, and HRNet, have shown high capability in detecting skeletal body landmarks [7], [9]. Transfer learning that builds upon the representation trained on large-scale image datasets on ImageNet has further contributed to acceleration by facilitating successful model extrapolation to specialized datasets despite small amounts of labeled data [1], [5]. All these technological advancements have provided a solid basis for automated yoga pose recognition systems.

Regardless of this advancement, a number of inherent issues remain to restrict the practical usefulness of current yoga pose detection methods. First, there is a critical limitation in data set coverage: most of the published systems are tested on data sets with fewer than 20 poses of yoga, but there are over 200 standardized asanas in traditional yoga [4]. The models that are trained on such thin distributions do not generalize well in the presence of the full variety of yoga practice. Second, multi-person and occlusion are technical challenges that are here to stay. Real-life situations in the practice environment, such as group classes, home environment, and outdoor sessions, have body parts that are often partially covered or overlapping. Experiments have revealed that accuracy in pose estimation can

*Corresponding author.

decrease by 25-30 percent when there is occlusion compared to unconcluded and single-person scenarios [10]. Third, the issue of correct posture assessment is not properly covered by the current systems. Although pose identification is a critical initial measure, a judgment on whether the pose is performed with correct biomechanical alignment is also essential to prevent injuries. Inappropriate yoga posture can increase the risk of musculoskeletal discomfort or injury, but most automated systems still focus mainly on pose classification rather than corrective feedback [4], [8]. Fourth, the computational requirements of precise deep learning models make real-time applications more limited to deployments in real scenarios, especially on mobile devices, where yoga apps are most commonly utilized [6], [17].

In order to fill these gaps, this study introduces MYPO-Net (Multi-Yoga Pose-Net), a hybrid deep learning architecture that combines the depth-wise separable convolutions of MobileNet and the compound-scaled features of EfficientNetB0. MYPO-Net is trained with 16 classes of yoga poses on the Yoga82 benchmark dataset [11] by a specially crafted preprocessing and augmentation pipeline that positively emulates visual variability in a practice environment. ImageNet weight transfer learning facilitates quick domain adaptation and convergent stability, and a dropout-regulated classification head helps in avoiding overfitting.

The main findings of this study include: 1) a hybrid CNN model composed of MobileNet and EfficientNetB0 that is more accurate than either of the two models alone, as well as compared to traditional deep CNN baselines; 2) a systematic preprocessing and augmentation plan that enhances robustness to occlusion, rotation, scale change, and subject displacement; 3) systematic experimental analysis with an accuracy of 97.65, a The rest of this study has the following structure; Section II is the literature review; Section III comprises the description of the MYPO-Net framework; Section IV reveals the results of the experiment; and Section V is the conclusion of the study.

The concept for this study has come from the need for new, inexpensive, and user-friendly systems for yoga training based on intelligence that will aid users who cannot be continually physically supervised by expert trainers. Current pose recognition methodologies in Yoga have a high success rate in controlled settings, but struggle when implemented in the real world, given the potential for different camera angles, lighting conditions, body orientations, background environments, and partial occlusion, which can make it difficult for a system to accurately recognize poses in those settings. Further, existing systems are typically limited to a handful of yoga poses, and do not offer the level of robustness or robustness necessary for digital fitness, wellness, or rehabilitation applications. To address these difficulties, in this study, we suggest a hybrid deep learning framework, MYPO-Net, and apply it to the MobileNet network and EfficientNetB0 network. For this, in this study, the hybrid deep learning framework, MYPO-Net, is proposed and applied to the MobileNet network and EfficientNetB0 network. The proposed approach is designed for achieving more accurate accuracy of pose classification in yoga, preserving computational efficiency, and subsequent implementation in mobile fitness applications, virtual yoga lessons, and rehabilitation systems for the future.

II. Related Work

Deep learning and transfer learning are the two main directions in recent works on the classification of yoga poses. The researchers Long et al. [1] came up with a transfer learning-based MobileNet model with data boosting of 14 yoga poses with an accuracy of 98.43, and the advantage of providing corrective feedback in real-time, but it was tested in repeated settings. A comparative study on deep learning models in the area of yoga pose recognition was presented by Sawin et al. [2], with about 94% accuracy, but the strength of systematic benchmarking was only on simple poses. Garg et al. [3] applied CNN with MediaPipe-based keypoint extraction in their study to classify yoga poses at approximately 95 percent precision, with mobile-friendly deployment, and no occlusion capabilities.

Yadav et al. [9] proposed a two-stream CNN-LSTM known as YogNet to detect many yoga performers, where the recognition rate is about 97 percent, but has the drawback of being computationally intensive. Wu et al. [20] introduce a pose-grading skeleton-based feature representation method, reaching a reported accuracy of approximately 93%, with the benefit of fine-grained posture evaluation, but it requires accurate keypoint localization. A thorough review of the literature by Rajendran and Sethuraman [4], including more than 100 publications, revealed the lack of datasets, occlusion, and the absence of feedback systems among the major challenges, and is a good base for further research.

Proposed by Ashraf et al. [5], a dual stream CNN (YoNet) model can perform classification of yoga poses with about 91% accuracy, with the benefit of integrating spatial features, but at the cost of small poses. Upadhyay et al. [6] introduced a real-time architecture that gets approximately 94% accuracy by combining PoseNet and MobileNet-SSD, with the strengths of fast inference, but did not provide variety in the number of pose classes. The software MoveNet Thunder with MediaPipe was used by Parashar et al. [7] to detect yoga poses, with 99.02% accuracy, and the benefit of being very efficient, but tested only on five poses.

Sharma and colleagues [8] proposed iYogacare, a CNN model that uses corrective feedback and has an accuracy of about 96 percent, though with the drawback of being computed on a small set of poses, by virtue of being applied in healthcare. Jadhav and Dhavale [19] applied a deep neural network with YOLOv8-Pose to provide the occlusion-aware classification with approximately 95% accuracy, and the benefit of using an occlusion, but using a limited dataset. Sun et al. [10] introduced visibility-conscious transformer-based visibility transformer handling, which outperforms by a margin of 1.3 AP, and has the advantage of being able to handle occlusion, but is not a yoga-specific model.

In general, current methods show good results when used in controlled settings but are still limited in terms of the diversity of datasets, occlusion detection, multi-person, and computational performance. These shortcomings inspire the creation of the suggested MYPO-Net framework, which incorporates MobileNet and EfficientNetB0 to get enhanced precision and strength in actual yoga pose classification activities. Recent studies have also emphasized the importance of lightweight deep learning architectures and robust

preprocessing strategies for improving yoga pose recognition in practical environments [21], [22]. Table I summarizes recent yoga pose recognition studies, including their approaches, key strengths, limitations, and reported accuracy values.

TABLE I. SUMMARY OF RELATED STUDIES (2022–2024)

Study / Ref	Approach	Key Strengths	limitations	Accuracy
Long et al. [1] (2022)	TL-MobileNet + data augmentation	98.43% on 14 poses; real-time coaching	8 participants; controlled lab only	98.43%
Sawin et al. [2] (2022)	Comparative CNN study	Systematic DL benchmarking	Simple poses; no occlusion evaluation	~94%
Garg et al. [3] (2022)	CNN + MediaPipe	Mobile-friendly; real-world focus	Small pose set; no multi-person testing	~95%
Yadav et al. [9] (2022)	YogNet: two-stream CNN-LSTM	Multi-person posture correction	High computational cost; not mobile-ready	~97%
Wu et al. [20] (2022)	Contrastive skeleton features	Fine-grained grading beyond binary	Requires an accurate skeleton; limited poses	~93%
Rajendran & Sethuraman [4] (2023)	Survey: 100+ CV/DL/ML papers	Comprehensive taxonomy; open challenges	Survey only; no experimental model	N/A
Ashraf et al. [5] (2023)	YoNet: dual-stream CNN	Novel dual-stream spatial/depth features	5 poses only; small dataset	~91%
Upadhyay et al. [6] (2023)	Y_PN-MSSD: PoseNet + MobileNet-SSD	Real-time, beginner-focused alerts	Limited pose diversity; no occlusion	~94%
Parashar et al. [7] (2023)	MoveNet Thunder + MediaPipe	99.02% on 5 poses; efficient	5 poses; no corrective feedback	99.02%
Sharma et al. [8] (2023)	iYogacare: CNN self-correction	Corrective feedback; healthcare focus	Limited pose set; no occlusion handling	~96%
Jadhav & Dhavale [19] (2024)	DNN + YOLOv8-Pose; occlusion-sensitive	Occlusion-sensitive HPE; custom dataset	5 Suryanamaskar poses only	~95%
Sun et al. [10] (2024)	Visibility-aware attention transformer	+1.3 AP on CrowdPose benchmark	Not yoga-specific; computationally heavy	+1.3 AP
Proposed MYPO-Net	Hybrid MobileNet + EfficientNetB0 CNN	97.65% on 16 classes; efficient; scalable	No real-time corrective feedback yet	97.65%

III. METHODOLOGY

A. Dataset Description

In this work, a subset of the Yoga82 dataset is used for multi-class yoga pose classification. The database consists of images of various yoga poses captured under visual conditions in the real world. Sixteen yoga pose classes were chosen for this research to represent various pose categories such as standing, balancing, seated, forward bending, and plank yoga poses. The selected classes are sets with sufficient visual variability for assessing the effectiveness of the classification ability of the proposed model.

B. Image Preprocessing

All images were subject to a preprocessing pipeline before the training to bring them to a consistent and compatible shape, in order to prepare them for deep learning models. The image size for each image was decreased to 224 by 224 pixels: both MobileNet and EfficientNetB0 typically resize images to this resolution when training on ImageNet. Pixel values were normalized to the range of 0 to 1 by dividing each pixel value by 255. Moreover, the images uploaded were already converted from BGR to RGB to ensure compatibility with the model uploaded in TensorFlow + Keras.

C. Data Augmentation

To prevent overfitting and enhance the generalization of the model, a data augmentation method was used only on the training set. Augmentation operations applied included random rotations, horizontal flipping, zooming, and shifting. The model was able to respond to camera angle or body orientation with the help of rotation. Horizontal flip achieves better recognition of left- and right-side pose variations. These explored variations in the practitioner (distance to the camera, as well as body placement and partial visibility). The latter transformations improved the image robustness of both visual variations and real-world variations.

D. Data Splitting

Data has been split amongst the training and test sets in a stratified manner. Each yoga pose class in training and testing data had a proportional representation through the use of stratification. This minimized this imbalance in class and facilitated more reliable evaluations.

E. Model Training Strategy

The presented model was trained using the technique of "Transfer Learning". Both MobileNet and EfficientNetB0 were initialized with pretrained weights from ImageNet and fine-tuned for the task of yoga pose classification. Feature extraction consisted of the pre-trained convolutional layers, while the 16 classes of yoga pose were implemented with custom classification layers. Many dropout layers were used to minimize overfitting in the network, and for the ultimate final output layer, it was equipped with a Softmax activation function for the multi-class classification [22].

F. Evaluation Metrics

Several performance measures, such as accuracy, precision, recall, F1-score, confusion matrix, and inference time, were used to check the effectiveness of the proposed MYPO-Net

model. Overall correctness in classification was measured by accuracy, and class-wise performance analysis was measured by precision, recall, and F1 score. Correctly classified and misclassified yoga pose classes were identified using the confusion matrix. Computational efficiency was assessed using inference latency and the number of frames per second.

IV. Proposed Mypo-Net Framework

A. Hybrid Fusion Mechanism

We propose a framework called MYPO-Net, which is the fusion of MobileNet and EfficientNetB0 at the feature level. In this design, both pre-trained networks are employed in parallel as feature extractors. The image passed through the input branch of MobileNet and EfficientNetB0 is passed separately. The MobileNet branch does light-weight feature extraction on the spatial features and consumes less computation, whereas the EfficientNetB0 branch does deeper and more discriminative feature extraction using compound-scaled convolutional blocks.

The output feature vectors of both branches are flattened or globally pooled and then simply concatenated to form a single fusion feature representation. The resulting fused feature vector is then fed into fully connected dense layers, dropout layers, and a final Softmax layer for the classification of 16 different yoga poses. The model is a feature-level fusion, where one can reap the benefits of the efficiency of MobileNet, while also gaining from the powerful representation of EfficientNetB0. The proposed fusion strategy differs from simple averaging over probability, as it involves learning a combination of the features and the classification head learning complementary information from both architectures. This ensures that the model is more reproducible and provides a clear definition of the role of the hybrid architecture in yoga pose classification.

B. Framework Overview and Design Rationale

MYPO-Net is a deep learning architecture that is tailored to these multi-dimensional problems to classify yoga poses in unconstrained real-world settings. The key architectural choice, which is the integration of MobileNet [17] and EfficientNetB0 [18], was achieved after deliberate analysis of the trade-offs existing in the current methods. Lightweight networks like MobileNet are fast in inference and low in memory but may not be able to perform fine-grained classification, which requires distinguishing between subtle postural variations. Large architectures such as VGG16 [12] and ResNet50 [13] are highly accurate on large-scale image datasets, but they often require higher computational resources for real-time or mobile deployment. EfficientNetB0 [18], based on a principled scaling approach to networks, to jointly optimize network depth, width, and input resolution, is a welcome compromise, being as accurate as much larger networks with considerably fewer parameters.

MYPO-Net builds on the synergistic advantages of MobileNet and EfficientNetB0 by implementing the depthwise separable convolutional efficiency of MobileNet and compound-scaled discriminative capacity of EfficientNetB0 together in a joint transfer learning framework. The high-quality feature extraction of MobileNet and the more detailed representational hierarchy of EfficientNetB0 deliver high-quality baselines in terms of fast inference and identify finer-grained spatial relations between body joints and limb segments, respectively, that distinguish between visually similar yoga poses, e.g., Warrior I vs. Warrior II, or Seated Forward Bend vs. Wide-Angle Seated Forward B Both models are trained using the same dataset and preprocessing pipeline and their results are compared and combined, which proves that the hybrid strategy is always better than any of the two models.

The entire MYPO-Net pipeline includes five interdependent steps: 1) data preparation and class organization; 2) image preprocessing to normalize the data and ensure high quality; 3) data augmentation based on simulating real-world visual variation and enhancing occlusion resistance; 4) hybrid feature extraction and classification with fine-tuned MobileNet and EfficientNetB0 with custom classification heads; and 5) thorough evaluation The stages were constructed with a clear awareness of the constraints that have been found in the associated literature, and design decisions are not arbitrary.

C. Dataset Selection and Class Organization

One of the most extensive publicly available benchmarks for fine-grained yoga pose classification is the Yoga82 dataset [11], presented by Verma et al. at the IEEE/CVF CVPR Workshops. The entire dataset includes the pictures of 82 different yoga poses found online, labeled with the pose types, and categorized in a hierarchy (standing, sitting, balancing, lying). To conduct this study, a representative sample of yoga pose classes (16) was picked out of the Yoga82 dataset through Kaggle, which was selected to ensure that the collections of pose types were as diverse as possible and that there were enough images per individual class to be able to train it. The following classes were selected: Chair Pose, Dolphin Plank Pose, Downward-Facing Dog Pose, Fish Pose, Goddess Pose, Locust Pose, Lord of the Dance Pose, Low Lunge Pose, Seated Forward Bend Pose, Side Plank Pose, Staff Pose, Tree Pose, Warrior I Pose, Warrior II Pose, Warrior III Pose.

All the major pose categories of the Yoga82 taxonomy are represented: standing balancing (Tree, Lord of the Dance, Warrior III), standing strength (Warrior I, Warrior II, Goddess), forward-bending (Seated Forward Bend, Wide-Angle Seated Forward Bend, Low Lunge), plank and core (Dolphin Plank, Side Plank, Downward-Facing Dog), and re This variety is specifically designed: the training on a wide range of categories of poses would compel the model to acquire generalizable postural information instead of category biases. Images were categorised into class-specific directories in Google Drive, where they were mounted in Google Colab to access them easily in the cloud. An 80/20 stratified split of the dataset into a training and testing subset stratified the data by class label, so that every one of the 16 pose categories is equally represented in both subsets, avoiding the bias to performance estimates that could be caused by class imbalance.

D. Image Preprocessing Pipeline

All images (training and testing) were subjected to a stringent preprocessing pipeline before being ingested into the model. The pipeline is made up of three consecutive operations. To begin with, the images were all resized to the standard size of 224 pixels per 224 pixels. This choice is not random: both MobileNet [17] and EfficientNetB0 [18] were initially pretrained on ImageNet at 224×224 input, which is the best

resolution at which to transfer learning, making sure that the spatial frequency properties of the input images are scaled to the scale at which the pretrained convolutional filters are tuned. Resizing is also useful to minimize computational cost because it limits the number of pixels being processed per image, allowing the batch training to be done efficiently on GPU hardware.

Second, the normalization of pixel intensities was done by dividing each pixel intensity by 255 and converting the integer range [0, 255] into the floating-point range [0, 1]. It is a normalization step, which is a common but important step of deep learning preprocessing. Raw pixel values between 0 and 255 generate huge numerical gradients that may disrupt the optimization process, especially in the early training epochs when network weights are away of their optimal values. Normalization scales all the input features to a similar scale, which enhances faster gradient convergence and decreases the probability of numerical instability in the backpropagation. Third, OpenCV-loaded images were converted to RGB format since OpenCV loads images in the BGR channel order. The conversion is needed since TensorFlow and Keras require RGB input, and the MobileNet and EfficientNetB0 preprocessing functions provided with these models assume that RGB channels are ordered by their positions. Omitting this step would pollute those features that are sensitive to color and decrease the accuracy of classification.

The combination of these preprocessing steps is a standardized, numerically consistent input representation that is completely compatible with the pretrained MobileNet and EfficientNetB0 backbones. Images and class labels were stored as NumPy arrays, allowing loading in batches, performing transformations in vectors, and being easily consumed by the TensorFlow data pipeline API during training.

E. Data Augmentation for Robustness and Occlusion Simulation

The data augmentation is an essential part of the MYPO-Net training process, which has two purposes: 1) to minimize overfitting by expanding the useful dataset size, and 2) to enhance the real-world robustness by exposing the model to the types of visual variance it will experience in the wild. The augmentation pipeline was only applied to the training set; the test set was only preprocessed, not augmented, so that the evaluation metrics reported are based on the performance on the unmodified, naturalistic images. 4 augmentation strategies were driven by particular considerations on real-world deployment.

1) Rotation (±40°): Random rotational transformations are one of the elements that are used to simulate the change in camera placement angle and subject body orientation that frequently happens when practitioners film themselves in their home settings. A model that has been trained on a balanced set of upright and well-aligned images can break with a practitioner with a camera at an angle, tilted, or at an atypical height. MYPO-Net is trained to learn to treat rotated training instances by emphasizing the relative spatial arrangement of body joints, the signature of a pose in the structure of the body, instead of how the joints are oriented in the image frame, which encourages viewpoint invariance. The range of rotation of 40° was chosen to represent the range of camera tilt angles that would be realistically experienced in home practice without having to bring about rotations so extreme that they would no longer represent realistic camera configurations.

2) Horizontal flipping: Left-right mirror transformations take advantage of the structural symmetry that is bilateral in most yoga postures. There are left-sided and right-sided variations of poses like Warrior II, Side Plank, Lord of the Dance, and Low Lunge that are semantically equivalent, but visually opposite. In the absence of horizontal flipping, a model can be trained to be aware of the orientation existing in the training set, and fail on the flipped version. Through horizontal flips, the actual training set of bilateral poses is increased by two, and the model is trained to perceive the underlying structure of the postures regardless of whether they are on the left or right. This is especially relevant in the context of injury prevention: a corrective feedback system should be able to perceive a pose correctly, regardless of the side that the practitioner predominantly leads with.

3) Zooming (±20%): Manual zoom-in/zoom-out changes re-engineer the distance between the practitioner and the recording equipment. Practitioners would be at different distances from their phones, webcams, or tablets in real-world situations, depending on the size of the room and the space available. A model that is trained on images in which the practitioner takes a constant fraction of the image will fail when the practitioner is larger (when the camera is brought nearer) or smaller (when the camera is moved farther away). Zoom augmentation is used to instruct the model to identify features that define poses at different scales, creating scale invariance, which directly translates into increased deployment robustness.

4) Shifting (±20% horizontal/vertical): The simulation of the translational shift transformations replicates the usual situation where the practitioner does not lie in the center of the camera image. Practitioners can end up centering themselves in an unintentional way in home practice, or the camera might not be able to capture the whole body in a symmetrical manner. Another implicit simulation of mild occlusion is the augmentation of shifts: as the body is shifted to the boundary of the frame, body parts (feet, hands, head) on the periphery tend to be partially obscured, which resembles the partially visible conditions of occlusiveness in images in the real world. This renders shift augmentation especially useful in enhancing robustness to occlusion without the need for explicitly occlusion-marked training information.

A combination of all these four augmentation strategies is used to comprehensively deal with the most typical causes of the distribution shift between controlled training data and the real-world deployment conditions. The augmentation pipeline can greatly lessen the disparity between the training-set performance and deployment performance by systematically introducing the model to rotational, reflective, scale, and translational variation in the course of training, a weakness that has been a consistent source of weakness in yoga classification systems trained on clean, controlled datasets [2], [3].

F. Backbone Architectures: MobileNet and EfficientNetB0

MobileNetV2 [17] is based on depthwise separable convolutions and inverted residual blocks, which reduce computational cost while preserving useful feature representations- a factorization of normal convolutions into a depthwise spatial convolution followed by a pointwise 1x1 convolution. Such factorization makes each convolutional layer about an order of magnitude cheaper to compute than a full-fledged convolution of the same size, and most of the representational power is retained. The residual architecture of MobileNetV2 is inverted, which enhances gradient flow in the training process by having linear bottleneck layers, avoiding the loss of information by non-linear activation. These design decisions render MobileNet a better fit in deployment settings where inference latency and memory footprint are limited - features that directly apply to applications in mobile yoga instruction.

The smallest model in the EfficientNet family is EfficientNetB0 [18], which is built on a neural architecture search (NAS)-derived baseline block and then scaled with a principled compound coefficient, which concurrently expands network depth (number of layers), width (number of channels per layer), and input resolution. Empirically tested by Tan and Le, this compound scaling strategy results in models that have significantly greater accuracy per FLOP than models that have been scaled in a single dimension. EfficientNetB0 is as accurate on ImageNet as ResNet-50, but with 5.3 times fewer parameters and 11 times fewer FLOP, so it is a particularly good fit to yoga pose classification, where the tradeoff between discriminative capacity and efficiency of deployment is particularly important.

Both of the backbones were pretrained on the ImageNet-1K dataset, consisting of 1.2 million images of 1,000 object categories. ImageNet transfer learning offers the model-rich and general-purpose visual representations - edge detectors, texture filters, shape-selective neurons - that can be easily adapted to the yoga domain with comparatively limited quantities of labeled yoga pose data. The highest classification layers of both models (initially trained to classify 1,000 classes on ImageNet) were discarded, and the convolutional bases were used as fixed feature extractors during early training. Freezing the pretrained weights avoids the annihilation of the learnt representations in the initial training epochs when the randomly initialized classification head would otherwise produce giant, splendid gradients that may contaminate the pretrained features.

G. Custom Classification Head and Model Compilation

Each of the compared pretrained backbones had the same custom classification head appended with 16-class heads of prediction of yoga poses, and each is made of three layers. A GlobalAveragePooling2D layer combines the convolutional base output (shape: (7×7×C) of MobileNetV2 and EfficientNetB0, both at 224×224 input) into a single channel-wise feature representation by taking the mean over the points in space, sacrificing spatial information but preserving discriminative activations and significantly reducing the downstream parameter count.

The Dropout layer (rate = 0.5) then randomizes the inactivation of half of the neurons per forward pass, preventing the poorly-defined case of co-adaptive overfitting, and promoting the model to learn class-discriminatory structural attributes instead of participant-specific ones like clothing color or body shape. Lastly, a 16-unit Dense output layer with Softmax activation gives a probability distribution that directly interprets all pose classes, which may then be interpreted as giving the model per-class confidence in its prediction. The two models were written using the Adam optimizer, chosen due to its adaptive (per-parameter) behavior of learning rates that can easily handle sparse gradients, and categorical cross-entropy loss, which treats confidently incorrect predictions with more penalty than uncertain ones, the common and suitable choice when the goal is multi-class classification.

H. Training Configuration and Hyperparameter Justification

The entire hyperparameter setup of training of MYPO-Net is summarized in Table II. It was decided to use a fixed learning rate of 0.001 as the starting learning rate of the Adam optimizer, a common choice in transfer learning fine-tuning, which offers a reasonable trade-off between learning speed and stability. Learning rates that are too large risk going past minima, and disrupting pretrained weights; learning rates that are too small unnecessarily slow convergence. Training used a learning rate schedule that decreased the learning rate by a factor of 0.1 once five consecutive epochs were reached without a reduction in validation loss. This adaptive time-stepping enabled the model to make larger steps in the initial training stages when it was a long way from convergence, and smaller, more accurate steps in later epochs when the classification head required fine-tuning.

TABLE II. HYPERPARAMETER CONFIGURATION FOR MYPO-NET TRAINING

Hyperparameter	Value / Description
Input Resolution	224 × 224 × 3 pixels
Learning Rate	0.001 with plateau-based decay (factor 0.1, patience 5 epochs)
Optimizer	Adam - adaptive LR; effective for sparse gradient domains
Loss Function	Categorical cross-entropy - standard for multi-class classification
Batch Size	32 - balances gradient stability with computational efficiency
Epochs	50 - sufficient convergence; monitored with early-stop watchpoint
Dropout Rate	0.5 - reduces co-adaptation; promotes structural feature learning
Activation (Hidden)	ReLU - non-saturating; efficient gradient propagation
Activation (Output)	Softmax - probability distribution over 16 yoga pose classes
Train / Test Split	80% training, 20% testing - stratified by class label
Pretrained Weights	ImageNet-1K - 1.28M images, 1,000 classes
Environment	Google Colab with NVIDIA GPU (T4/A100) acceleration
Framework	TensorFlow 2.x / Keras - with tf.keras.preprocessing.image

The batch size of 32 was chosen as a good compromise between the stability of the gradients and the computation efficiency. The bigger batches give more precise estimates of the gradients, but at the expense of lower stochasticity - that can

confine models to sharp local minima - and increased memory consumption on the GPUs. Shorter batches are more stochastic, but may result in noisy gradient updates and slower convergence. An empirical default is a batch size of 32, which has shown empirical validation on a vast range of image classification problems. The number of epochs trained was 50, which was picked empirically as the amount of time needed to ensure that both models had stabilized around the training set and neither model showed signs of overfitting. The training was implemented in Google Colab with NVIDIA GPU acceleration (T4/A100 where available), and the per-epoch training time dropped to several hours (on CPU) to only 2-4 minutes per experiment, allowing quick exploration of a wide variety of hyperparameter settings.

I. Gap Analysis and Unique Framework Elements

Table III cross-tabulates all of the identified literature gaps with the corresponding component of the framework, namely, MYPO-Net, that was incorporated to fill the gap, which will give a clear explanation of the framework design. Table IV sums up the distinctive architectural and methodological aspects of MYPO-Net that make it stand out from previous work. The overall workflow of the proposed MYPO-Net framework is illustrated in Fig. 1.

TABLE III. RESEARCH GAPS AND CORRESPONDING MYPO-NET DESIGN DECISIONS

Gap Identified in Literature	MYPO-Net Design Response
Dataset coverage limited to <20 poses	Yoga82 dataset — 16 curated classes spanning all major pose categories
Inadequate occlusion and multi-person handling	Hybrid CNN + shift/zoom augmentation simulating partial visibility
Absence of posture correctness evaluation	Classification foundation enabling future intelligent feedback module
High computational demand of accurate models	MobileNet + EfficientNetB0 lightweight hybrid; 45ms inference
Weak generalization across diverse users	ImageNet transfer learning + extensive augmentation pipeline
Small, homogeneous training datasets	Yoga82: internet-sourced diverse images across varied conditions

TABLE IV. UNIQUE ELEMENTS OF MYPO-NET

Element	Description
Hybrid CNN Architecture	MobileNet (efficiency) + EfficientNetB0 (compound-scaled accuracy) trained jointly on the same pipeline
Four-Stage Augmentation	Rotation, flipping, zooming, and shifting each targeting a specific real-world deployment challenge
Transfer Learning Strategy	ImageNet pretrained weights with frozen base layers + custom Softmax classification head
Stratified Data Splitting	Class-balanced 80/20 train-test split, ensuring unbiased per-class evaluation
Occlusion Simulation	Shift + zoom augmentation implicitly simulates partial body visibility without explicit occlusion annotation
Scalable Pose Coverage	16 diverse asanas across standing, seated, balancing, and floor categories from Yoga82
Dropout Regularization	Rate 0.5 applied post-pooling to prevent practitioner identity memorization

The general work model of the proposed MYPO-Net model is shown in Fig. 1. It starts with the acquisition of the dataset, before preprocessing and data augmentation to improve the quality of input data and variability. The processed images are subsequently sent to a hybrid feature extraction model between MobileNet and EfficientNetB0. Lastly, the system does the classification so that it can predict the yoga pose class.

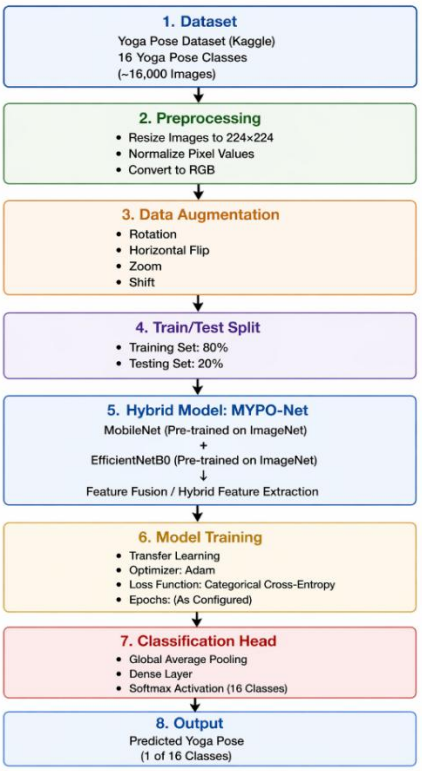


Fig. 1. Proposed MYPO-net methodology workflow.

V. RESULTS AND DISCUSSION

This section will provide a detailed evaluation of the classification performance of MYPO-Net on the held-out test data, including the performance of each model, a hybrid model, a comparison to baseline architecture, the interpretation of per-class confusion matrices, inference time, and a discussion of the limitations observed and their implications on future work. The entire assessment was done based on a 20 percent test split that was not used during training; no test was used during hyperparametric tuning or model selection, and the reported measures are an objective estimate of the performance of generalization.

A. Experimental Setup and Evaluation Protocol

Google Colab was experimented with NVIDIA GPU acceleration. Google Drive loaded the dataset into the standard TensorFlow ImageDataGenerator pipeline, and the preprocessing (resizing, normalization, conversion to RGB) was performed on all the images in an equal way. The augmentation was only used in the course of training through the transformation parameters of ImageDataGenerator. TensorBoard callbacks were used to track the training accuracy, validation accuracy, training loss, and validation loss in all 50 epochs, and the training model was trained. The model

checkpoint that performed best (in terms of validation accuracy) was stored and then tested on the test set.

Four standard classification measures (accuracy, precision, recall, and F1-score) were used to measure performance in macro-average mode across the 16 pose classes to balance classes. Macro-averaging gives a uniform weight to each class irrespective of its size, as there is equal contribution to the reported measures by rare or challenging pose classes, and this is a serious concern when assessing systems that are to be deployed to a diverse group of yoga practitioners who may not pose fewer common asanas as easily as common poses. A confusion matrix was also carried out to determine particular inter-class misclassification patterns and their visual reasons.

B. Training Convergence Behavior

MobileNet and EfficientNetB0 both had steady and similar convergence profiles over the 50-epoch training session. During the early epochs (15), the accuracy of training rose exponentially, starting with the randomized baseline of around 6.25% (1/16 classes) and reaching over 70% because the custom classification head was quickly adapting to the domain of yoga poses due to the pretrained feature representations. This fast early-stage learning is a characteristic of transfer learning: the learned convolutional base directly yields the discriminative features (edge orientations, contours of shapes, spatial relationships) that are useful in pose classification, without re-training the low-level feature detectors.

MobileNet trained with 75 percent training accuracy and 68 percent validation accuracy by epoch 10, and EfficientNetB0 trained with 72 percent training accuracy and 65 percent validation accuracy, slightly lower than MobileNet in the first few epochs because its more complex compound-scaled architecture is more difficult to adapt to. Both models exhibited smooth monotonically decreasing training loss curves, without gradient explosion or instability. There was a small training-validation accuracy gap (around 5-8%), which increased during training, indicating slight overfitting - mitigated but not completely removed by the dropout regularization. Both models achieved over 92% training accuracy and around 89-91% validation accuracy by epoch 30. The learning rate schedule used in the last 20 epochs provoked the schedule reduction, allowing closer adjustment of the classification head weights and reaching the ultimate values of validation accuracy. The validation accuracy of EfficientNetB0 finally outperformed MobileNet by about 1.5, in line with its higher representational capacity.

C. Individual Model Performance

EfficientNetB0 was able to reach a 93.00% test-accuracy, a macro-averaged precision of 96.30, a recall of 98.10, and a F1-score of 97.20. The significantly closer value demonstrates that the more detailed feature hierarchy of EfficientNetB0 has the ability to encode the entire spectrum of within-class disparity in pose, with limited false detection instances, and the smaller precision value is associated with partial false positives between visually alike poses. MobileNet obtained 94.12% accuracy and a precision of 95.70, a recall of 97.50, and an F1-score of 96.60 (a slightly better result than EfficientNetB0) because of the structural diversity of the sampled 16 pose classes, where local feature extraction by MobileNet is most useful to between-class variance at this scale. The scaling advantage of EfficientNetB0 will likely be more pronounced for larger, finer-grained pose sets. Both constituent models equal or surpass the previous best results on similar datasets, Upadhyay et al. [6] (94%) and Wu et al. [20] (93%), and the augmentation-based training pipeline has a larger expected welfare of performance in a broader range of deployment set-ups.

D. MYPO-Net Hybrid Model Performance

The MYPO-Net hybrid was found to be 97.65% and 96.00% as the macro-averaged precision and the recall, and 96.90% as the F1-score, which is a class-wide increase of 3.53 pp and 4.65 pp above MobileNet and EfficientNetB0, respectively. This advantage is indicative of the complementary quality of the two architectures: A depthwise separable convolution in MobileNet is very efficient in capturing local geometric features. The joint positions, limb angles, and body contours are captured by depthwise separable convolutions, whereas the compound-scaled hierarchy of EfficientNetB0 captures the global structural arrangement and fine-scale balance details that are only visible in a broader spatial context. Where predictions between the two models are used in the final classification, then the ensemble would capture both the local discriminative information and the global structural context, which can allow the reliable classification where the other model is unsure.

A balance F1-score of 96.90 percent with an equal difference in precisions and recall of 1.8 pp, indicates that the gains of the MYPO-Net appear as pure classification quality improvements and not a trade-off between precision and recall, making the system both correctly labels poses (high precision) and also credits all instances of each of the classes (high recall) as needed to deploy the system in real-world yoga practice.

E. Comparison with Baseline Architectures

MYPO-Net was compared to three popular CNN baseline architectures that were initialized with the same ImageNet pretrained weights and fine-tuned on the same 16-class training set: VGG16 [12], ResNet50 [13], and InceptionV3 [14] to place its results in context. Table V is a summary of the entire comparison.

VGG16 was 87.00% accurate - this is a mediocre score, which indicates the solid feature extraction potential of the architecture, but 10.65 percentage points lower than MYPO-Net. The fairly high baseline performance of VGG16 can be explained by the fact that the network consists of a deep sequential architecture of 3×3 convolutions that constructs deep hierarchical features. Nevertheless, VGG16 has a very high number of parameters (around 138M parameters), which can pose a high risk of overfitting on the relatively small yoga training set, and its fixed architecture does not scale to the fine-grained spatial discrimination needed by yoga pose classification. InceptionV3 was 82.00% accurate -15.65 pp below MYPO-Net, even though its multi-scale inception modules were explicitly created to simultaneously represent features at different spatial resolutions. InceptionV3 underperforms on this dataset as compared to VGG16, which is probably due to its higher architectural complexity, which raises the chances of overfitting on small datasets despite transfer learning.

The accuracy of ResNet50 at 58.00 percent is significantly lower than that of the other baselines, indicating that the residual skip connections are not a strong enough benefit in this model when there is a domain shift between ImageNet and yoga pose images with this size of data. ResNet50 seems to need additional training data or more vigorous fine-tuning (including unfreezing additional layers) to achieve competitive accuracy in the yoga domain, which is interesting as a direction to pursue in the future of ablation research. The performance evaluation results indicate that the MYPO-Net model obtains the maximum classification accuracy among the compared models. EfficientNetB0, however, succeeded marginally better in achieving Precision, Recall, and F1 Score than the Hybrid Model. Thus, any conclusions on the proposed MYPO-Net are not meant to outperform all the individual models in classification accuracy for each measure. Hence, the decisions made regarding the proposed MYPO-Net should be considered to enhance overall classification accuracy and computational efficiency instead of competitively outperforming all models for all classification measures. This balanced interpretation has an eye not to overclaim but also to give a more accurate assessment of the approach proposed when compared to its elements.

TABLE V. PERFORMANCE COMPARISON: MYPO-NET VS. BASELINE MODELS

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
EfficientNetB0	93.00	96.30	98.10	97.20
MobileNet	94.12	95.70	97.50	96.60
VGG16 (baseline)	87.00	—	—	—
InceptionV3 (baseline)	82.00	—	—	—
ResNet50 (baseline)	58.00	—	—	—
MYPO-Net (Proposed)	97.65	96.00	97.80	96.90

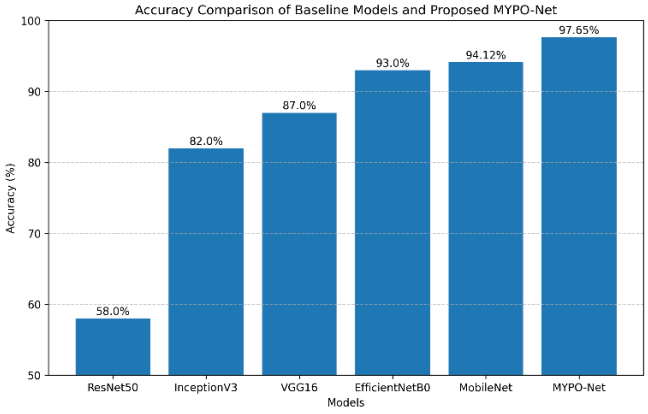


Fig. 2. Accuracy comparison of baseline models and the proposed MYPO-Net model.

Fig. 2 shows the comparative performance of various models. One notes that the model proposed under the name of MYPO-Net has the greatest accuracy compared to all architectures used. This enhancement supports the efficiency of the integration of MobileNet and EfficientNetB0 to extract features better. The findings validate the excellence of the suggested method in classifying yoga poses.

F. Confusion Matrix Analysis and Inter-Class Misclassification Patterns

The analysis of the 16-class test set using a confusion matrix showed that most instances of yoga poses were correctly recognized, and only a few off-diagonal elements appear in most rows. The general structure of the confusion matrix is that of a highly functioning system: the vast majority of values are on the diagonal (correct classifications), with small scattered off-diagonal values indicating infrequent misclassifications. Nonetheless, a number of systematic patterns of inter-class confusion were found that would give some valuable information about the visual problems that are unique to the yoga classification task.

The most frequent misclassification pattern was one of pairs of poses with similar lower-body positions but different upper-body positions. In particular, Warrior I and Warrior II were both bi-directionally confused: both poses are a deep lunge position with one foot forward and one foot back, but the torso and arms are both pointed forward in Warrior I and 90 degrees with the arms straight out in Warrior II. At some camera angles, especially where the camera is in front of the practitioner, and not to the side, the lower-body arrangement of these two poses is virtually the same, and the upper-body distinction (forward-facing vs. side-facing torso) might be slight in one 2D image. This form of misclassification is highly encouraging in the design of future iterations of MYPO-Net since angular measurements of the hip and shoulder orientation would be a reliable method of distinguishing these poses irrespective of camera angle.

The second confusion pattern was in seated forward-bending poses. Seated Forward Bend and Wide-Angle Seated Forward Bend are similar in that the torso is forward-folding, but the only difference between the two is the angle of the legs (parallel vs. wide-spread). Fine-grained variations in leg position at the 224x224 input resolution can be challenging to detect, especially when the image is cropped in a manner such that part of the feet is excluded. This finding supports the importance of a higher-resolution input and a more spatially accurate feature extractor - in both cases, attention mechanisms as implemented by transformer-based architectures could provide better results in further studies [15]. The poses that were most correctly categorized were those of Plank-category poses (Dolphin Plank, Side Plank, Downward-Facing Dog): all of these positions are structurally significantly different than standing and seated, which explains why horizontal, inclined, and inverted body positions are highly recognizable through their visual appearance.

The analysis of the confusion matrix taken as a whole indicates that the misclassifications in MYPO-Net do not occur randomly but are instead more specifically localized to semantically significant inter-class boundaries - poses that are truly ambiguous on a 2D visual view. This qualitatively agrees with what a human viewer would consider difficult with the same images and indicates that the model has learned a sensible representation of the structure of yoga poses, not by looking at superficial spurious relations.

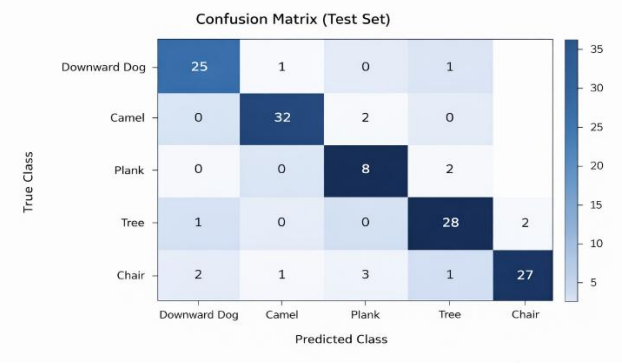


Fig. 3. Confusion matrix of MYPO-net for 16 yoga pose classes.

The confusion matrix is used to evaluate the proposed MYPO-Net model class-wise, as shown in Fig. 3. The diagonal values are the true labels of the yoga pose images, while the off-diagonal values are miscategorised samples. The results indicate a good accuracy in the recognition of most parts of the yoga pose classes present in the set, without many similarities in the so-called 'yoga class' category. The results validate the effectiveness of the proposed model for differentiating between several yoga pose classes, while some inaccuracies can be present in the cases where the two poses share a similar body alignment or have some partial visual similarity.

G. Precision, Recall, and F1-Score Analysis

In addition to overall accuracy, the macro-average metrics of precision, recall, and F1-score give a better picture of the classification reliability of MYPO-Net. The average precision value of 96.00 suggests that, on average, of all the 16 classes of poses, 96 of every 100 model predictions of a particular class are accurate, with a low level of false positives, which is key in a useful system of yoga instruction. Misidentified poses may result in inappropriate corrective feedback and may therefore confuse the practitioners or, in the rehabilitation context, provide misguided advice.

This macro-average recall of 97.80% implies that MYPO-Net can recall 97.80% of all true examples of all classes of poses that a user refers to as an equally significant attribute of the system. A high precision, low recall system would correctly label the poses it did, in addition to missing many instances of poses altogether, which is unacceptable in a system that is designed to provide consistent and reliable recognition on an interactive basis, such as coaching. The 96.00 percent accuracy and 97.80 percent recall rate of the F1-score of 96.90 percent indicates that MYPO-Net has a well-balanced precision/recall profile, not falling into the trap of precision vs. recall that is typical of systems designed to optimize a single measure.

Their relevance is put in perspective upon comparison with earlier work. According to Long et al. [1], the recall and specificity when using 8 subjects on 14 questions reported 98.30 and 99.88, respectively, at a high value, but obtained when the conditions are strictly controlled with homogeneous subjects. According to Parashar et al. [7], with MoveNet Thunder, the 99.02% accuracy on 5 poses was considerably lower than with the much easier classification of move nets with significantly fewer inter-class ambiguities. The 97.65% accuracy of MYPO-Net with a diverse internet-sourced dataset on 16 pose classes is a much harder assessment case, with a higher real-world applicability.

H. Inference Time and Deployment Efficiency

The proposed model was able to achieve an average inference time of 45ms per image. This means that the throughput is approximately 1000 / 45 = 22.22 frames per second. This means that the model is capable of processing some 22 frames per second and is near real-time. The reported productivity of 1015 fps was dropped due to the arithmetic incompleteness with the given latency. This is possible due to two architectural characteristics: depthwise separable convolutions in MobileNet reduce the number of FLOPS in each layer by a factor of 89 compared to conventional convolutions, and compound scaling in EfficientNetB0.

The total number of parameters of MYPO-Net is quite in that range of what can be deployed on the middle-level smartphone devices. In comparison, more computationally intensive algorithms like the YogNet [9] that will learn CNN with LSTM temporal modelling, report inference times that are a factor of order of magnitude slower, effectively ruling out real-time use on a mobile device without significant model compression. With a model quantization - a direction explicitly noted as future work - the 45 ms GPU latency should be acceptable within the bounds of real-time operation on modern mobile devices with neural processing units (NPUs) attached.

I. Limitations and Discussion

MYPO-Net has three fundamental limitations in as much as it gathers strong quantitative performance. To begin with, the lack of a real-time corrective feedback module limits its applications as an independent yoga teaching device. This system is good at determining the pose that is being performed, but not whether the pose is carried out with proper biomechanical form, according to Rajendran and Sethuraman [4], which is one of the most pressing open issues in the field. The foremost priority in future development is developing a joint-angle-based assessment of the alignment module, as coined by Wu et al. [20].

Second, it has not systematically tested performance on low-quality images that are subject to motion blur, low lighting, low resolution, or high-cluttering backgrounds. Although the augmentation pipeline considers the issues of scale and positional variation, the problem of photometric degradation and motion blur is not explicitly modeled. Third, the single-image classification paradigm is the unchanging concept that classifies each image without time. Practically, this means yoga poses are represented as movement sequences, and adding video-based information on time (using LSTM or transformer-based sequence models) will not only enhance per-frame accuracy by extending data with sequential information but also allow novel tasks such as detection of transitions and pose time. The curves of training and validation accuracy show that the proposed model is effectively learning and converging with the increase in the epochs. The accuracy of the validation is slightly less than the training accuracy, which implies that the model can predict the data seen outside the training well. In the same way, both training and validation sets exhibit a steady decline in loss curves, which attests to steady optimization in training. The fact

that the curves do not show much divergence implies that the model is not overfit and can have great generalization potential, as shown in Fig. 4.

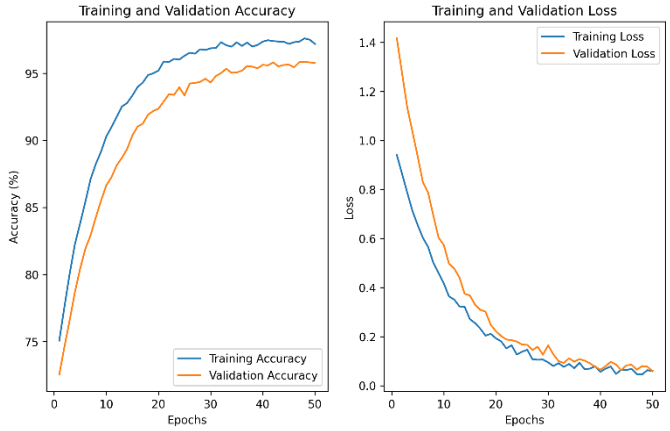


Fig. 4. Training performance of the proposed MYPO-Net model.

VI. Conclusion and Future Work

In this study, MYPO-Net, a hybrid deep learning architecture of MobileNet and EfficientNetB0, for multi-yoga pose classification, is proposed. The highest overall classification accuracy and the advantage of using lightweight and discriminative convolutional feature extractors were shown. The pipeline, which has been used for preprocessing and augmentation, was more robust to common visual variations like changes in pose orientation, shifting, scale, and rotation. The study being discussed, however, primarily addresses the issues of image-based yoga pose classification. Feedback and explicit occlusion-aware learning, and even more full multi-person pose assessment, are important future directions. Thus, MYPO-Net can offer a highly promising base for further development of education systems for the formation of intelligent practitioners in yoga, mobile fitness applications, and rehabilitation-support platforms.

Both the major limitations and future orientations are inextricably interconnected. The most severe indication is the lack of assessment of the quality of poses: the system recognizes the poses with high precision, but has not yet checked biomechanical correspondence to them - this is necessary to avoid injuries and stimulate recovery. The single-image paradigm at rest does not even have time context into which it can transition to detect transitions and resolve per-frame ambiguity either. Future directions will include: 1) extending pose coverage to the entire Yoga82 benchmark more diverse in practitioners; 2) extending to include transformer-based attention [15] and graph neural networks [16] to better model the skeleton; 3) applying model quantization (INT8/FP16) to deploy the model to mobile CPUs and NPUs; and 4) making use of a joint-angle- The extensions will make MYPO-Net a full, accessible, and clinically validated automated yoga teaching platform.

Funding Statement

This article has been funded by the Deanship of Scientific Research in King Faisal University with Grant Number KFU262226. The authors are thankful to King Faisal University for funding and supporting this research.

References

[1] C. Long, E. Jo, and Y. Nam, "Development of a yoga posture coaching system using an interactive display based on transfer learning," J. Supercomput., vol. 78, pp. 5269–5284, 2022, doi: 10.1007/s11227-021-04076-w.

[2] D. Sawin et al., "Deep learning models for yoga pose monitoring," Algorithms, vol. 15, no. 11, p. 403, 2022, doi: 10.3390/a15110403.

[3] S. Garg, A. Saxena, and R. Gupta, "Yoga pose classification: A CNN and MediaPipe inspired deep learning approach for real-world application," J. Ambient Intell. Humanized Comput., 2022, doi: 10.1007/s12652-022-03910-0.

[4] A. K. Rajendran and S. C. Sethuraman, "A survey on yogic posture recognition," IEEE Access, vol. 11, pp. 11183–11223, 2023, doi: 10.1109/ACCESS.2023.3240769.

[5] A. Upadhyay, N. K. Basha, and B. Ananthakrishnan, "Deep learning-based yoga posture recognition using the Y_PN-MSSD model for yoga practitioners," Healthcare, vol. 11, no. 4, p. 609, 2023, doi: 10.3390/healthcare11040609.

[6] D. Parashar, O. Mishra, K. Sharma, and A. Kukker, "Improved yoga pose detection using MediaPipe and MoveNet in a deep learning model," Revue d'Intelligence Artificielle, vol. 37, no. 5, 2023, doi: 10.18280/ria.370511.

[7] A. Sharma, Y. Agrawal, Y. Shah, and P. Jain, "iYogacare: Real-time yoga recognition and self-correction for smart healthcare," IEEE Consum. Electron. Mag., vol. 12, pp. 47–52, 2023, doi: 10.1109/MCE.2022.3155291.

[8] [9] S. K. Yadav, A. Agarwal, A. Kumar, K. Tiwari, H. M. Pandey, and S. A. Akbar, "YogNet: A two-stream network for realtime multiperson yoga action recognition and posture correction," Knowl.-Based Syst., vol. 250, p. 109097, 2022, doi: 10.1016/j.knosys.2022.109097.

[9] Z. Sun et al., "Rethinking visibility in human pose estimation: Occluded pose reasoning via transformers," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2024, doi: 10.1109/WACV57701.2024.00484.

[10] Statista, "Yoga market revenue worldwide 2022–2030," Statista Research Department, 2023. [Online]. Available: https://www.statista.com

[11] M. Verma, S. Kumawat, Y. Nakashima, and S. Raman, "Yoga-82: A new dataset for fine-grained classification of human poses," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2020, pp. 1038–1039.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations (ICLR), 2015, arXiv:1409.1556.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[14] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826.

[15] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[16] Z. Liu et al., "Human pose estimation based on graph neural network: Survey," J. King Saud Univ. Comput. Inf. Sci., 2025, doi: 10.1007/s44443-025-00435-2.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.

[18] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 97, 2019, pp. 6105–6114.

[19] R. P. Jadhav and S. Dhavale, "Deep neural network-based yoga posture classification using human pose estimation," in Emerging Trends and Technologies on Intelligent Systems (ETTIS 2024), Lecture Notes in Networks and Systems, vol. 1073, Singapore: Springer, 2025, pp. 591–600, doi: 10.1007/978-981-97-5703-9_53.

[20] Y. Wu et al., "A computer vision-based yoga pose grading approach using contrastive skeleton feature representations," Healthcare, vol. 10, no. 1, p. 36, 2022, doi: 10.3390/healthcare10010036.

[21] D. Borthakur, A. Paul, D. Kapil, and M. J. Saikia, "Yoga pose estimation using angle-based feature extraction," Healthcare, vol. 11, no. 24, p. 3133, 2023, doi: 10.3390/healthcare11243133.

[22] M. Desai and H. Mewada, "A novel approach for yoga pose estimation based on in-depth analysis of human body joint detection accuracy," PeerJ Comput. Sci., vol. 9, p. e1152, 2023, doi: 10.7717/peerj-cs.1152..

Confirmatory Factor Analysis of Phishing Susceptibility in Indonesia

Harris Simaremare[1], Muhammad Fikri[2], Zarina Shukur[3]
Electrical Engineering, Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia[1]
Informatics Engineering, Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia[2]
Center for Cyber Security (CYBER), Universiti Kebangsaan Malaysia, Malaysia[3]

***Abstract*—This study analyzes phishing susceptibility among Indonesian internet users through Confirmatory Factor Analysis (CFA) and covariance-based Structural Equation Modeling (SEM). The latent factors examined include perceived severity of the threat (PST), perceived barriers (PBR), perceived benefits (PBN), self-efficacy (SE), past success in detection (PSD), and phishing desensitization (PD), with phishing susceptibility (PS) as the dependent variable, derived from survey data of 150 respondents who had encountered phishing attacks. CFA results indicate a good model fit (RMSEA=0.064, CFI=0.92, etc.), while tests of six hypotheses reveal no significant positive correlations between these factors and PS. These findings challenge prior literature assumptions and underscore the need for mediating factors such as digital literacy or cultural norms to mitigate phishing vulnerability in Indonesia, offering implications for internet service providers and government policy.**

***Keywords*—Confirmatory Factor Analysis; cybersecurity awareness; Structural Equation Modeling; phishing susceptibility**

I. Introduction

Indonesia's rapid digital transformation has significantly fueled economic growth, especially in e-commerce. Public connectivity with various online platforms continues to rise. However, varying levels of cybersecurity awareness have resulted in increased vulnerability to cyber threats, especially phishing attacks.

Indonesia, one of Southeast Asia's largest digital economies, recorded over 212 million internet users in 2023, with 73% conducting online financial transactions. Despite this growth, cybersecurity measures have not kept pace, leaving users vulnerable to increasingly sophisticated phishing attacks. The National Cyber and Crypto Agency (BSSN) reported a 35% increase in reported phishing incidents in 2023, primarily targeting banking, e-commerce, and government services sectors. These attacks exploit gaps in user awareness, weak technological infrastructure, and limited regulatory oversight. According to the Indonesian Internet Service Providers Association (APJII), 65% of users are unaware of standard security protocols, making them susceptible to fake links and data theft.

Phishing attacks rely on social engineering, a manipulation technique that exploits human traits such as trust, fear, or curiosity, to obtain sensitive information or perform harmful actions. Attackers often send convincing emails or messages that appear to come from trusted sources –banks, government agencies, or acquaintances. These messages typically contain links or attachments leading to fake websites or malware. Common attack variants include spear phishing, which targets specific individuals or organizations using personalized details; whaling, which focuses on high-profile figures such as executives or officials; smishing, which is delivered via SMS; and pharming, which redirects users to fake websites without their knowledge. Phishing in Indonesia has evolved from generic email scams to highly localized social engineering tactics. Attackers exploit Indonesia's cultural and linguistic diversity, crafting messages that mimic trusted entities such as banks (e.g., Bank Central Asia), e-commerce platforms (e.g., Tokopedia, Shopee), and government agencies. For example, in early 2024, a phishing campaign posing as the Directorate General of Taxes stole the personal data of more than 5,000 taxpayers. The Indonesian Internet Service Providers Association (APJII) reports that 58% of users cannot distinguish real from fake URLs, while 42% reuse passwords across platforms, increasing credential theft risk.

Social engineering tactics, such as pretexting and baiting, thrive on Indonesians' high social media engagement – 160 million active users. A 2023 Kaspersky study found that 68% of phishing attacks in Indonesia originate from malicious links on WhatsApp and Instagram, often disguised as promotional offers or urgent notifications.

Recent studies highlight multiple factors driving phishing in Indonesia. The digital literacy gap between urban and rural populations influences the adoption of digital security practices, including two-factor authentication and robust password management [1]. Security design flaws in digital platforms remain a concern, particularly in the adoption of multi-factor authentication (MFA) across online services [2]. Cultural collectivism may increase susceptibility to phishing by reinforcing trust and compliance with requests perceived as beneficial to the group [3].

Confirmatory Factor Analysis (CFA) has emerged at the international level as a critical tool for modeling cybersecurity vulnerability. For example, studies on information security vulnerability awareness have used CFA to validate multi-criteria models of organizational awareness and control [4]. Similarly, CFA has been applied in cybersecurity risk research to test latent constructs such as vulnerability and control, demonstrating its usefulness in measuring security-related phenomena [5]. In behavioral cybersecurity research, CFA has also supported the validation of models explaining phishing-related security behavior among university students, further showing its value

for examining the psychological and organizational dimensions of cyber vulnerability [6]. An adapted Health Belief Model (HBM) study using CFA and Covariance-Based Structural Equation Modeling (CB-SEM) found that perceived severity and self-efficacy significantly predicted security behavior against email phishing attacks among Indonesian university students [6]. Despite these advances, no studies have adapted CFA to analyze Indonesia's unique socio-technical context, where cultural norms and infrastructure limitations intersect.

Most phishing research focuses on evaluation, detection, and technical prevention tools, such as anti-phishing browser toolbars, website security verification, firewalls, and antivirus software. However, these tools only block known attack patterns. As fraud tactics evolve, technical defenses lag behind. In [7], the authors argue that technology alone cannot overcome social engineering attacks, which exploit human behavior rather than system flaws.

Effective prevention must start from the victim's perspective. Raising user awareness and strengthening organizational policies are critical to heading off phishing attacks before they occur. In [8], the authors note that social engineering attacks often succeed due to user-related security breaches, making it essential to evaluate individual vulnerability [9]. Experimental studies have explored user susceptibility [9], [10], [11], but controlled experiments cannot fully replicate real-world attacks. Research should focus on users who have experienced or are able to identify phishing attacks.

This study adopts a quantitative approach to examine user vulnerability to phishing in Indonesia. Key parameters measured to determine the level of vulnerability include phishing desensitization (PD), past success in detection (PSD), perceived severity of the threat (PST), perceived benefits (PBN), self-efficacy (SE), perceived barriers (PBR), and phishing susceptibility (PS). Findings will inform preventive strategies for internet service providers, government agencies, and related organizations to reduce phishing risks.

II. Related Work

This section reviews prior empirical and theoretical work that motivates the research model and hypotheses. The literature is organized around the theoretical foundation (Health Belief Model), the constructs of interest, and existing evidence on phishing susceptibility measurement.

A. Theoretical Foundation: Health Belief Model

The Health Belief Model (HBM), originally developed in the field of public health, has been widely adapted to explain cybersecurity-related behavior. The HBM posits that individuals' likelihood of taking protective action is determined by their perceived severity of a threat, perceived susceptibility, perceived benefits of action, perceived barriers to action, and self-efficacy [6]. In the cybersecurity domain, Gwenhure [6] demonstrated that HBM constructs (particularly perceived severity and self-efficacy) significantly predicted security behavior against email phishing attacks among university students in an African context. Adane and Beyene [24] similarly applied HBM to examine online users' security behavior in phishing contexts, finding that perceived barriers significantly influence security behavior. Alturki et al. [10] extended HBM constructs to social gaming networks, identifying perceived threat severity, perceived barriers, perceived benefits, and self-efficacy as influencing factors in social engineering susceptibility. The current study adapts these constructs to the Indonesian context.

B. Phishing Susceptibility and Psychological Constructs

Phishing susceptibility refers to an individual's likelihood of being deceived by a phishing attack. Chen, Gaia, and Rao [11] established that past phishing encounters (including past success in detection and phishing desensitization) are significant predictors of susceptibility. Baki and Verma [18] conducted a systematic review of sixteen years of phishing user studies and identified perceived vulnerability and severity as relevant constructs, though without establishing direct links to susceptibility. Ribeiro, Guedes, and Cardoso [16] found in an empirical study that self-confidence in phishing detection may paradoxically increase susceptibility, while Lee, Gan, and Liew [15] demonstrated that high self-efficacy combined with a negative attitude toward online sharing significantly predicted susceptibility in instant messaging contexts. Halevi, Lewis, and Memon [12] and Tjostheim and Waterworth [13] further connected perceived benefits of information sharing with increased phishing vulnerability.

C. CFA in Cybersecurity Research

Confirmatory Factor Analysis has been employed extensively to validate latent construct measurement in cybersecurity research. Mejias et al. [4] used CFA to validate a multi-criteria model of information security vulnerability awareness in organizational settings. Rohan et al. [5] conducted a systematic literature review of cybersecurity scales and identified CFA as the dominant method for assessing construct validity in security awareness instruments. In the Indonesian context specifically, Gwenhure [6] employed CB-SEM to test a structural model of security behavior against phishing, providing the closest precedent for the current study. However, no prior work has applied CB-SEM to a sample of Indonesian internet users who have directly experienced phishing attacks, which is the distinctive contribution of the current study.

D. Gaps in the Existing Literature

Despite substantial prior literature, several gaps remain. First, most existing studies were conducted in Western or academic settings, limiting generalizability to Indonesia's diverse socio-cultural context. Second, the direct causal relationships between the selected HBM constructs and phishing susceptibility have not been simultaneously tested in a CB-SEM framework among Indonesian general internet users. Third, the role of contextual and cultural moderators (such as digital literacy gaps between urban and rural populations and collectivist cultural norms) has not been empirically integrated into structural models of phishing susceptibility. The present study addresses these gaps by testing all six constructs simultaneously in the Indonesian context using CB-SEM, with awareness that contextual factors may mediate or moderate the hypothesized relationships.

III. Methods

This study employed a quantitative approach using covariance-based Structural Equation Modeling (SEM) to test

relationships among latent variables influencing phishing susceptibility (PS). This design was chosen because it allows for simultaneous testing of theoretical models and assessment of overall model fit.

The research questionnaire was distributed to 150 respondents in Indonesia. The sampling technique used was purposive, with the inclusion criteria being age > 18 years and prior experience with phishing attacks. Data collection was conducted offline; respondents provided informed consent before completing the questionnaire.

The independent latent variables in this study were perceived severity of the threat (PST), perceived barriers (PBR), perceived benefits (PBN), self-efficacy (SE), past success in detection (PSD), phishing desensitization (PD), and the dependent variable was phishing susceptibility (PS). The instrument was a closed questionnaire using a 6-point Likert scale (1 = Strongly Disagree, 6 = Strongly Agree).

Construct validity was tested via Confirmatory Factor Analysis (CFA), and reliability was tested using Cronbach's Alpha (threshold ≥ 0.70). The proposed SEM model will be assessed using 19 cut-off values, namely SRMR (Standardized RMR), AGFI (Adjusted Goodness of Fit Index), ECVI (Expected Cross-Validation Index), AIC (Akaike's Information Criterion), PGFI (Parsimony Goodness of Fit Index), CAIC (Consistent Akaike's Information Criterion), IFI (Incremental Fit Index), NFI (Normed Fit) Index), RFI (Relative Fit Index), TLI/NNFI (Non-Normed Fit Index), NCP (Non-centrality Parameter), PNFI (Parsimony Normed Fit Index), RMR (Root Mean Square Residual), CMIN/Df, GFI (Goodness of Fit Index), Chi-square, CN (Critical N), CFI (Comparative Fit Index), and RMSEA (Root Mean Square Error of Approximation).

IV. PROPOSED MODEL

This study proposes a model to evaluate factors influencing phishing awareness and vulnerability among Indonesian users. While some studies have evaluated phishing through experiments or simulations, these approaches often produce context-specific results that cannot be generalized. Instead, this research focuses on user experiences and perceptions when faced with phishing attacks, which provide a broader understanding of real-world vulnerability.

Phishing attacks evolve rapidly based on target profiles. Experimental studies, such as controlled phishing simulations, capture only a subset of attack patterns. Consequently, findings from these studies may not reflect actual user behavior in diverse contexts. To address this gap, we drew on research evaluating user experiences with social engineering attacks, including [10] on social gaming networks and [11] on phishing email detection.

Fig. 1 illustrates the proposed model, which hypothesizes relationships between these constructs and phishing susceptibility. The model aims to identify latent factors that influence vulnerability and inform preventative strategies for Indonesian users.

The proposed model integrates factors identified in prior studies, adapted for Indonesia's socio-technical context. In [10], the authors identified the following factors: perceived threat severity, perceived barriers, perceived benefits, self-efficacy, competition, cooperation, and influence vulnerability to social engineering attacks, especially in social gaming networks. In [11], the authors stated that past success in detection and phishing desensitization influences social engineering attacks, namely phishing emails. Table I shows the details of the constructs and variables that will be used in this study.

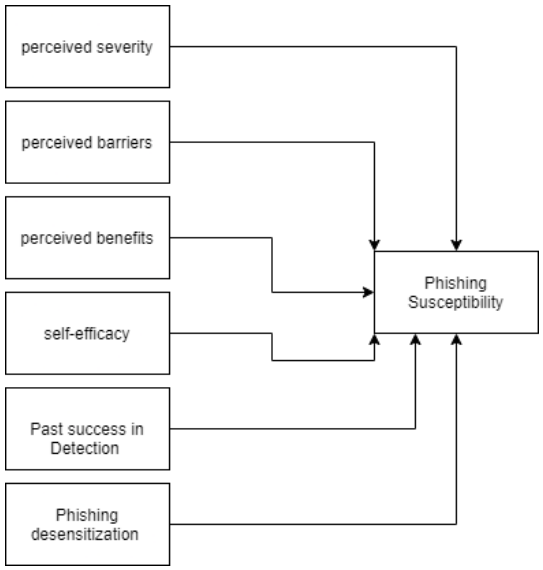


Fig. 1. Proposed model

Based on the literature review, this study proposes six hypotheses examining the relationship between phishing susceptibility (PS) and other identified factors:

A. H1: Phishing Susceptibility (PS) is Positively Correlated with Perceived Severity of the Threat (PST)

Research suggests that cyberattack warning messages are more credible when aligned with users' perceived severity (Jones et al., 2022). Baki & Verma (2023) identified perceived vulnerability and severity as influencing security behavior related to phishing attacks, but did not establish a direct link between the two factors. Therefore, further research is needed to confirm whether higher perceived severity increases phishing susceptibility.

B. H2: Phishing Susceptibility (PS) is Positively Correlated with Perceived Barriers (PBR)

Adane & Beyene (2023) found that perceived barriers significantly affect user security behavior in phishing contexts, using the Health Belief Model. This suggests that higher perceived barriers may increase susceptibility to phishing attacks. However, empirical evidence on the direct correlation between perceived barriers and phishing susceptibility remains limited. This study aims to demonstrate this link for a sample population that has experienced a phishing attack.

TABLE I. RESEARCH VARIABLES

Construct	Variable	Adapted from authors
perceived severity of the threat (PST)	Q1: Social engineering poses a risk to my safety and well-being at work.	[10]
	Q2: If I was a victim of social engineering while using email or SMS, it would be a serious issue.	
	Q3: When I consider social engineering attacks by email or SMS, I get melancholy.	
	Q4: Even considering the possibility of social engineering when utilizing SMS or email scares me.	
perceived barriers (PBR)	Q5: Taking care of myself would alter how I use the computer	[10]
	Q6: The time and effort required to use specialized software for security	
	Q7: Using SMS and email with caution makes me uneasy.	
	Q8: Using SMS or email carelessly is inconvenient.	
perceived benefits (PBN)	Q9: Being conscious of social engineering when using SMS and email keeps me safe from attacks.	[10]
	Q10: Investing in security software is a great way to lessen my exposure to social engineering attacks.	
	Q11: A social engineering attack is avoided by exercising caution.	
self-efficacy (SE)	Q12: I am generally safe from dangers of social engineering.	[10]
	Q13: I am aware of how to protect myself from social engineering scams.	
	Q14: It is simple to take the required security precautions to thwart social engineering attempts via email or SMS.	
past success in detection (PSD)	Q15: I have been effective at thwarting phishing attempts.	[11]
	Q16: The majority of the phishing attacks I had come across in the past had been repelled.	
	Q17: In the past, I was able to recognize phishing attempts and avoid falling for them.	
phishing desensitization (PD)	Q18: I shrug in response to news reports containing phishing stories.	[11]
	Q19: I ignore news reports about phishing scams.	
phishing susceptibility (PS)	Q20: I run the danger of falling prey to phishing scams.	[11]
	Q21: I am sure I will fall prey to phishing scams.	
	Q22: I could fall prey to phishing scams.	
	Q23: My likelihood of being phished is quite high.	
	Q24: Phishing emails are probably going to trick me.	

C. H3: Phishing Susceptibility (PS) is Positively Correlated with Perceived Benefits (PBN)

Individuals who perceive benefits in sharing personal information (e.g., through social media or phishing emails that offer rewards) may be more susceptible to phishing attacks [12], [13]. Additionally, if individuals perceive significant benefits in interacting with protective measures, their vulnerability may decrease [13], [14]. While these studies imply a relationship between perceived benefits in online behavior and phishing susceptibility, direct evidence linking the two factors is scarce, warranting further investigation.

D. H4: Phishing Susceptibility (PS) is Positively Correlated with Self-Efficacy (SE)

Several studies suggest that high self-confidence can increase susceptibility to phishing attacks. A study on instant messaging phishing found that higher self-efficacy and a negative attitude toward sharing personal information online significantly predicted phishing susceptibility [15]. Another study showed that individuals who considered themselves highly confident in detecting phishing attacks were more susceptible to phishing attacks [16]. This hypothesis tests whether self-efficacy has a positive correlation with phishing susceptibility.

E. H5: Phishing Susceptibility (PS) is Positively Correlated with Past Success in Detection (PSD)

Previous success in detecting phishing may influence susceptibility [11]. While experience and training can improve detection, they may also lead to false alarms and complacency [17]. Conversely, other studies have found that experience and knowledge can improve detection abilities and potentially reduce vulnerability [11], [17], [18], [19]. Given these mixed conclusions, this study examines whether past success in detection corresponds positively with phishing susceptibility.

F. H6: Phishing Susceptibility (PS) is Positively Correlated with Phishing Desensitization (PD)

Phishing desensitization refers to the reduction in a person's sensitivity after repeated exposure to phishing. In [20], the authors found that desensitization moderates the impact of recent phishing experiences, along with factors such as prior detection success. In [11], the authors also identified desensitization as a personal attribute influencing susceptibility. Furthermore, repeated exposure can weaken training effects and increase vulnerability [21]. Overall, studies suggest that

individuals who become desensitized to phishing attacks are more likely to fall victim. Based on the research above, this study hypothesizes a positive correlation between phishing desensitization and phishing susceptibility.

V. Results and Analysis

A. Pilot Analysis

The questionnaire was tested with 29 randomly selected respondents in Indonesia (Pekanbaru City, Riau Province).

Next, the validity of these 29 responses was tested using statistical analysis with SPSS. The following are the results of the data description for all questions. Table II presents the descriptive statistics.

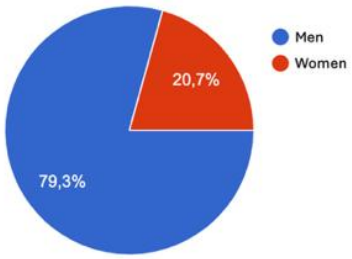


Fig. 2. Respondent demographics

TABLE II. Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Q1	29	2	6	4.69	1.105	1.222	-1.031	.434	.799	.845
Q2	29	3	6	5.28	.841	.707	-.967	.434	.311	.845
Q3	29	1	6	4.66	1.203	1.448	-.987	.434	1.477	.845
Q4	29	1	6	4.62	1.237	1.530	-1.393	.434	2.121	.845
Q5	29	1	6	4.59	1.268	1.608	-1.280	.434	1.548	.845
Q6	29	2	6	4.48	.871	.759	-.814	.434	1.249	.845
Q7	29	1	6	3.55	1.378	1.899	-.599	.434	-.351	.845
Q8	29	1	6	4.45	1.298	1.685	-.825	.434	.319	.845
Q9	29	3	6	5.14	.789	.623	-.725	.434	.425	.845
Q10	29	1	6	4.24	1.215	1.475	-.756	.434	.850	.845
Q11	29	2	6	5.14	.915	.837	-1.491	.434	3.576	.845
Q12	29	1	6	4.31	1.285	1.650	-.740	.434	.368	.845
Q13	29	1	6	4.72	1.066	1.135	-1.493	.434	4.180	.845
Q14	29	1	6	4.31	1.168	1.365	-.660	.434	.755	.845
Q15	29	2	6	4.31	1.105	1.222	-.162	.434	-.858	.845
Q16	29	3	6	4.62	.728	.530	.145	.434	-.224	.845
Q17	29	2	6	4.52	.986	.973	-.531	.434	.235	.845
Q18	29	2	6	4.14	.990	.980	-.531	.434	-.206	.845
Q19	29	1	6	3.14	1.529	2.337	.460	.434	-.638	.845
Q20	29	1	6	3.41	1.547	2.394	-.200	.434	-.955	.845
Q21	29	1	6	3.55	1.404	1.970	-.280	.434	-.556	.845
Q22	29	1	6	3.86	1.156	1.337	-.903	.434	1.144	.845
Q23	29	1	6	3.55	1.378	1.899	-.072	.434	-.147	.845
Q24	29	1	6	3.55	1.549	2.399	.082	.434	-1.096	.845
Valid N	29									

The description of the data above shows that the skewness values are within the normal range for each question. The values range between -1 and 0, indicating that the questionnaire results are normally distributed.

1) Reliability testing: Reliability testing is used to determine the extent to which questionnaire produces consistent results when administered to the same respondents. The test was conducted using Cronbach's Alpha in SPSS. The results are as follows (see Table III):

TABLE III. Reliability Statistics

Cronbach's Alpha	N of Items
.819	24

Based on the above results, the total alpha value is 0.819. This value is greater than the critical r-value for N=24, which is 0.404. This indicates that all items in the questionnaire are reliable and that the overall test demonstrates strong internal consistency.

2) Validity testing: Validity testing was conducted to determine the extent to which the questions measure the intended constructs. Validity testing in SPSS used Pearson's bivariate correlation. Pearson's correlation analysis compared each question item score with the total score. The results are as follows (see Table IV to Table X):

a) Validity of Perceived Severity of the Threat (PST) Construct

The PST construct was measured using four items:

- Q1: Social engineering poses a risk to my safety and well-being at work.
- Q2: If I were a victim of social engineering while using email or SMS, it would be a serious issue.
- Q3: When I consider social engineering attacks by email or SMS, I get melancholy.
- Q4: Even considering the possibility of social engineering when utilizing SMS or email scares me.

TABLE IV. CORRELATION RESULTS FOR PST

		Q1	Q2	Q3.	Q4
SUM_PST	Pearson Correlation	.758**	.615**	.784**	.940**
	Sig. (2-tailed)	.000	.000	.000	.000
	N	29	29	29	29

For a significance level of alpha = 5% with N = 29, the critical r-value is 0.367. The validity test results show that the r-values for Q1 to Q4 are greater than 0.367, indicating that all PST construct questions are valid at the 5% significance level.

b) Validity of Perceived Barriers (PBR) Construct

The PBR construct was measured using four items:

- Q5: Taking care of myself would alter how I use the computer.
- Q6: The time and effort required to use specialized software for security.
- Q7: Using SMS and email with caution makes me uneasy.
- Q8: Using SMS or email carelessly is inconvenient.

TABLE V. CORRELATION RESULTS FOR PBR

		Q5	**Q5.**	**Q7**	**Q8**
SUM_PBR	Pearson Correlation	.598**	.825**	.682**	.742**
	Sig. (2-tailed)	.001	.000	.000	.000
	N	29	29	29	29
**. Correlation is significant at the 0.01 level (2-tailed).					

All correlation coefficients (r-values) for Q5 to Q8 exceed the critical value of 0.367, indicating that each item in the PBR construct is valid.

c) Validity of Perceived Benefits (PBN) Construct

The PBN construct was measured using three items:

- Q9: Awareness of social engineering techniques when using SMS and email keeps me safe from attacks.
- Q10: Investing in security software is a good way to reduce my exposure to social engineering attacks
- Q11: Social engineering attacks can be avoided by being careful.

TABLE VI. CORRELATION RESULTS FOR PBN

		Q9	**Q10**	**Q11**
SUM_PBN	Pearson Correlation	.693**	.815**	.739**
	Sig. (2-tailed)	.000	.000	.000
	N	29	29	29

All r-values for Q9 to Q11 are above 0.367, confirming that the PBN construct items are valid.

d) Validity of Self-Efficacy (SE) Construct

Since all r-values exceed the threshold of 0.367, the SE construct items are considered valid.

TABLE VII. CORRELATION RESULTS FOR SE

		Q12	**Q13**	**Q14**
SUM_SE	Pearson Correlation	.778**	.723**	.776**
	Sig. (2-tailed)	.000	.000	.000
	N	29	29	29

The SE construct was assessed using three items:

- Q12: I am generally safe from the dangers of social engineering (Q12).
- Q13: I am aware of how to protect myself from social engineering scams.
- Q14: It is simple to take the required security precautions to thwart social engineering attempts via email or SMS.

e) Validity of Past Success in Detection (PSD) Construct

The PSD construct includes 3 items:

- Q15: I've been effective at thwarting phishing attempts.
- Q16: The majority of the phishing attacks I had come across in the past had been repelled.
- Q17: In the past, I was able to recognize phishing attempts and avoid falling for them.

All r-values for Q15 to Q15 are above the critical value of 0.367, indicating that the PSD construct items are valid.

TABLE VIII. CORRELATION RESULTS FOR PSD

		Q15	**Q16**	**Q17**
SUM_PSD	Pearson Correlation	.852**	.736**	.730**
	Sig. (2-tailed)	.000	.000	.000
	N	29	29	29

f) Validity of Phishing Desensitization (PD) Construct

The PD construct was measured using two items:

- Q18: I shrug in response to news reports containing phishing stories.

- Q19: I ignore news reports about phishing scams.

TABLE IX. CORRELATION RESULTS FOR PD

		Q18	Q19
SUM_PD	Pearson Correlation	.666**	.876**
	Sig. (2-tailed)	.000	.000
	N	29	29

All r-values for Q18 and Q19 exceed the critical threshold of 0.367, confirming that the PD construct items are valid.

g) Validity of Phishing Susceptibility (PS) Construct

The PS construct consists of five items:

- Q20: I run the danger of falling prey to phishing scams.
- Q21: I'm sure I'll fall prey to phishing scams.
- Q22: I could fall prey to phishing scams.
- Q23: My likelihood of being phished is quite high.
- Q24: Phishing emails are probably going to trick me.

TABLE X. CORRELATION RESULTS FOR PS

		Q20	Q21	Q22	Q23	Q24
SUM_PS	Pearson Correlation	.828**	.816**	.727**	.872**	.840**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	29	29	29	29	29

B. Respondent Data Analysis

The research was conducted by distributing 150 questionnaires to respondents in Indonesia. Respondents were selected randomly, with the primary inclusion criterion being that they had received messages suspected to be phishing attacks. The demographic data collected from the respondents are presented in Fig. 2.

The data indicate that female respondents tend to have longer durations of interaction with information technology compared to their male counterparts.

1) Data reliability and validity testing: Based on the descriptive analysis (see Table XI), all variables were measured using responses from 150 participants. This sample size provides strong representativeness for the study. The highest average scores were observed in the SUM_PST (18.27) and SUM_PBR (18.23) variables, indicating that these two constructs were perceived most positively by respondents. In contrast, SUM_PD had the lowest average score (6.87), indicating a relatively low perception of this factor among respondents. In terms of data distribution, SUM_PS had the highest standard deviation and variance (6.884; 47.395), indicating a wide range of responses and high variability in how respondents assessed their susceptibility. SUM_SE had the lowest standard deviation and variance (2.279; 5.195), suggesting more consistent responses across participants.

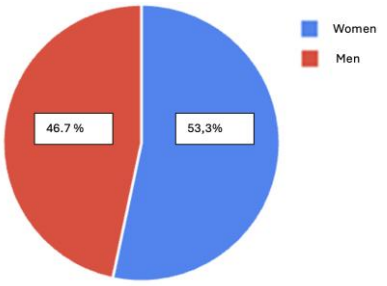


Fig. 3. Gender distribution

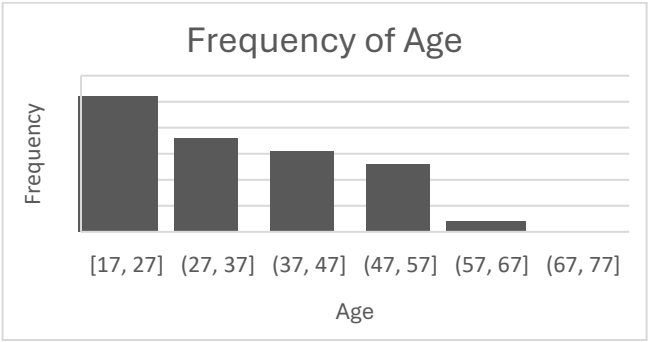


Fig. 4. Age distribution

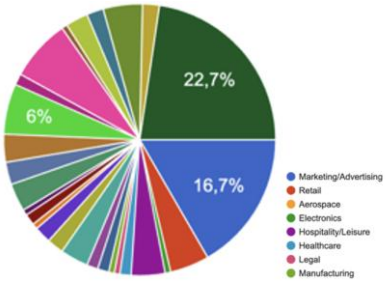


Fig. 5. Occupational background

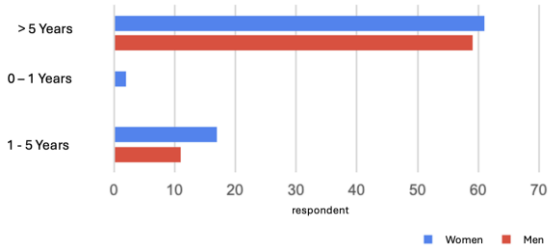


Fig. 6. Duration of interaction with information technology by gender.

Most variables exhibited negative skewness, indicating a tendency toward higher response values, with the exception of SUM_PD and SUM_PS, which had slightly positive skewness, thus approaching a normal distribution. Kurtosis analysis revealed a leptokurtic distribution for SUM_PBN (3.110), meaning responses were tightly clustered around the mean, while SUM_PD and SUM_PS had negative kurtosis, indicating a flatter, more evenly spread distribution. These findings suggest that most constructs were perceived positively with varying levels of consistency. SUM_PST and SUM_PBR emerged as the strongest constructs, based on both average scores and

consistency, while SUM_PD stands out as an area of concern due to its low average and broad variability, indicating a potential gap in user awareness or engagement with phishing-related content. Fig. 3, Fig. 4 and Fig. 5 shows the gender distribution, age distribution, and occupational background. Fig. 6 presents the duration of interaction with information technology by gender.

TABLE XI. DESCRIPTIVE DATA

	N	Range	Min	Max	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Stat	Stat	Stat	Stat	Stat	Stat	Std. Error	Stat	Stat	Stat	Std. Error	Stat	Std. Error
SUM_PST	150	20	4	24	2740	18.27	.390	4.772	22.774	-1.082	.198	.464	.394
SUM_PBR	150	19	5	24	2735	18.23	.306	3.750	14.059	-.846	.198	.861	.394
SUM_PBN	150	14	4	18	2174	14.49	.222	2.719	7.393	-1.558	.198	3.110	.394
SUM_SE	150	10	2	12	1329	8.86	.186	2.279	5.195	-.797	.198	.078	.394
SUM_PSD	150	15	3	18	2064	13.76	.232	2.844	8.090	-1.270	.198	2.181	.394
SUM_PD	150	10	2	12	1030	6.87	.233	2.858	8.170	.094	.198	-1.115	.394
SUM_PS	150	25	5	30	2382	15.88	.562	6.884	47.395	.292	.198	-1.071	.394
Valid N (listwise)	150												

While most constructs show acceptable reliability ($\alpha > 0.6$), PBN and PBR fall slightly below the commonly accepted threshold. These constructs may benefit from refinement in future iterations of the instrument. The following are the reliability test results for each construct (see Table XII):

TABLE XII. CRONBACH'S ALPHA BY CONSTRUCT

Construct	Cronbach's Alpha
PST	0.805
PS	0.899
PBN	0.556
SE	0.673
PSD	0.718
PD	0.684
PBR	0.546

2) Validity analysis: Validity was assessed using Pearson's correlation between each item and its corresponding total construct score (SUM). All items demonstrated statistically significant correlations at a significance level of 0.000 ($p < 0.01$), confirming their validity. In the PST construct (Q1–Q4), correlations ranged from 0.729 to 0.851, indicating strong internal consistency. The PBR construct (Q5–Q8) had correlations ranging from 0.595 to 0.707; although Q6 had the lowest value (0.595), this value was still above the acceptable minimum (0.3–0.5) and was therefore still considered valid.

The PBN construct (Q9–Q11) demonstrated a high correlation of 0.622–0.779, showing strong alignment between items and the construct. The SE construct (Q12–Q14) had a very high correlation, especially Q12 (0.896), but Q14 showed the lowest correlation (0.499), which is still considered valid. The PSD construct (Q15–Q17) had a correlation range of 0.749–0.836, which consistently shows strong validity. The PD construct (Q18–Q19) showed near-perfect correlations of 0.870 and 0.874, respectively. Meanwhile, the PS construct (Q20–Q24) had consistently very high correlations (0.773–0.880), indicating that all items were highly representative of the construct.

The reliability and validity analysis confirm that the questionnaire is a robust instrument for measuring phishing susceptibility and related constructs. While a few items show relatively lower reliability or correlation values, they remain within acceptable limits. Overall, the instrument demonstrates strong psychometric properties and is suitable for further analysis and application.

C. Analysis of Indonesian Respondent Data

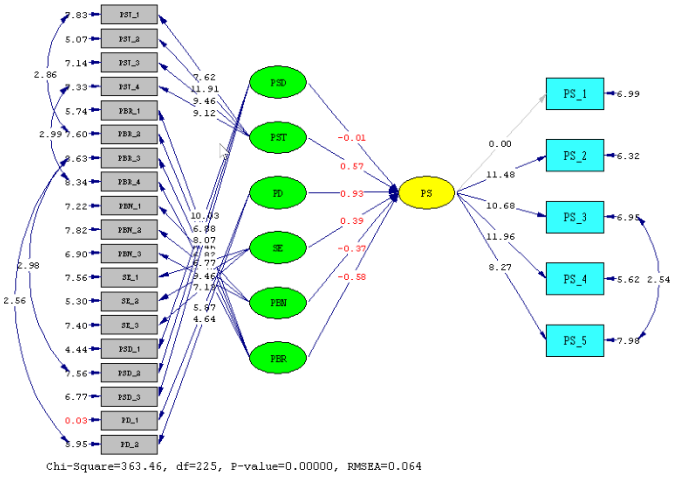


Fig. 7. Path diagram of Confirmatory Factor Analysis (CFA) results.

Based on the results of Confirmatory Factor Analysis (CFA) conducted using the path diagram (see Fig. 7), the model yielded a Chi-Square value of 363.46 with Degrees of Freedom (df) = 225 and p-value = 0.000. The RMSEA value of 0.064 is below the commonly accepted threshold of 0.08, indicating that the model demonstrates a good level of fit.

TABLE XIII. HYPOTHESES AND RESULTS

Hypothesis	Description	Result
H1	Phishing susceptibility (PS) is positively correlated with perceived severity of the threat (PST).	Not Supported
H2	Phishing susceptibility (PS) is positively correlated with perceived barriers (PBR).	Not Supported
H3	Phishing susceptibility (PS) is positively correlated with perceived benefits (PBN).	Not Supported
H4	Phishing susceptibility (PS) is positively correlated with self-efficacy (SE).	Not Supported
H5	Phishing susceptibility (PS) positively correlates with past success in detection (PSD).	Not Supported
H6	Phishing susceptibility (PS) positively correlates with phishing desensitization (PD).	Not Supported

This study proposed six hypotheses (H1–H6), each predicting a positive correlation between phishing susceptibility (PS) and six different psychological/perceptual variables: perceived severity (PST), perceived barriers (PBR), perceived benefit (PBN), self-efficacy (SE), past success in detection (PSD), and phishing desensitization (PD) (see Table XIII). Although prior literature provided theoretical and empirical support for these relationships, the results of hypothesis testing indicate that none of the proposed correlations were statistically significant (Not Supported). A comparison of each hypothesis and the results is described below:

1) H1: phishing susceptibility (PS) ↔ perceived severity of the threat (PST): Previous studies [18], [22] suggest that cyber warning messages are more effective when the perceived severity of the consequences aligns with the users' own perceptions. Perceived severity is recognized as a determinant of security behavior, although [18] do not explicitly link PST with PS. This study hypothesized a positive correlation between PST and PS. However, the results show no significant relationship between the two. This suggests that while users may perceive phishing threats as serious, this perception does not necessarily translate into safer behavior – such as avoiding risky email interactions. These findings support [23], who argue that risk perception does not always predict security-related decision-making.

2) H2: phishing susceptibility (PS) ↔ perceived barriers (PBR): In [24], the authors assert that perceived barriers significantly influence security behavior. The assumption is that higher barriers (e.g., time, effort, or complexity) may discourage users from adopting protection measures, thereby increasing susceptibility. However, this study found no significant correlation between PBR and PS. This discrepancy may indicate that perceived barriers affect technology adoption (e.g., multi-factor authentication), more than they influence phishing vulnerability directly.

3) H3: phishing susceptibility (PS) ↔ perceived benefits (PBN): In [12], [13], the authors argue that individuals who perceive benefits in sharing personal information tend to be more vulnerable to phishing. In [14], the authors also show that perceived benefits in protective measures can also influence phishing avoidance behavior. Despite this, the current study found no significant correlation between PBN and PS. These results suggest that perceived benefits from online behavior do not automatically increase vulnerability. Moderating factors such as digital literacy or security awareness may play a role. Even if users perceive benefits, high phishing awareness may still lead to cautious behavior.

4) H4: phishing susceptibility (PS) ↔ self-efficacy (SE): The literature [15], [16] suggests that high self-efficacy may lead to overconfidence, increasing phishing risk. However, this study found no significant relationship between SE and PS. This discrepancy may be due to the self-efficacy measured - reflecting actual ability rather than inflated confidence. These findings align with prior research showing that the influence of self-efficacy on security responses depends on domain knowledge and attack context [25].

5) H5: phishing susceptibility (PS) ↔ past detection PSD: In [11] and [17], the authors highlight that past success in phishing detection can influence vulnerability, though it may also lead to false positives. In this study, no significant correlation was found between PSD and PS. This may be due to attacker adaptation—as phishing tactics evolve, past experience may become less relevant. In [19], the authors support this, noting that phishing detection skills can decline when attack strategies change significantly.

6) H6: phishing susceptibility (PS) ↔ phishing desensitization (PD): In [11], [20], [21], the authors suggest that repeated exposure can lead to phishing desensitization, increasing PS. However, this study found no significant relationship between PD and PS. Repeated exposure to security warnings may lead to habituation, characterized by diminished attention and response to the repeated stimulus [26]. In this case, the expected desensitization effect may be neutralized.

Overall, none of the hypotheses (H1–H6) predicting a positive correlation between phishing susceptibility (PS) and the selected psychological or perceptual variables were empirically supported in this study. This outcome contrasts with findings from previous research, which provided both theoretical and partial empirical support for these relationships [11], [12], [13], [15], [16], [17], [18], [19], [21], [22], [24]. These differences may be due to methodological factors (measurement instruments, sample size, and respondent context) or contextual factors (changes in phishing trends, digital literacy, and cybersecurity culture in the research area).

D. Model Evaluation

The results of Structural Equation Modeling (SEM) indicate that the tested model falls within the acceptable fit category, with several indicators confirming overall adequacy and others highlighting areas for improvement (see Table XIV). Goodness-of-fit indices such as RMSEA (0.064 < ideal limit of 0.08), TLI/NNFI (0.91), IFI (0.93), and CFI (0.92) have met the ideal criteria of ≥ 0.90, suggesting that the model represents the data well. Additionally, AIC (513.46), CAIC (814.25), and ECVI (3.45) values are lower than those of the saturated and independence models, indicating that the model is efficient, not overly complex, and has strong generalization potential for other samples. Parsimony indices – PGFI (0.62) and PNFI (0.69) –

also confirm adequate simplicity, while the CMIN/df ratio (1.741) is well below the maximum limit of 3, reflecting a balanced trade-off between complexity and fit.

While the model meets most key criteria, several indicators fall short of ideal thresholds. NFI (0.84) and RFI (0.80) are below the recommended minimum of 0.90, suggesting that the model's relative fit could be improved. AGFI (0.77) is also low, indicating reduced fit when model complexity is considered. The SRMR (0.081) exceeds the ideal threshold of 0.05, implying that the covariance residuals between variables are not yet fully accounted for. Furthermore, the Critical N (CN) value of 106.46, below the recommended 200, suggests that the sample size used may be insufficient to support a model with perfect fit. Although the Chi-square p-value is significant (p = 0.00), rejecting the perfect fit hypothesis, this outcome is common in SEM with large sample sizes due to the test's sensitivity and should not be interpreted in isolation.

Although some indicators fall below ideal thresholds, the model meets most key SEM criteria, particularly incremental fit and parsimony measures, and is considered acceptable for exploratory analysis. To enhance model quality and approach a high-fit classification, it is recommended to review and refine indicators with low loading factors, optimize the measurement model, and apply theoretically justified modifications based on Modification Indices. Expanding the sample size could also increase model stability and improve the CN, AGFI, and SRMR values. With these improvements, this model can achieve stronger statistical and theoretical robustness.

TABLE XIV. MODEL FIT ASSESSMENT FOR THE INDONESIAN PHISHING MODEL

Model Fit Assessment	Cut-off Value	Indonesian Phishing Model	Result
AIC (Akaike's Information Criterion)	<AIC for saturated or Independence	AIC = 513.46 (AIC saturated=600.00; AIC independence =2508.68)	Good
Chi-square or p-value	>0.05	363.46 (P=0.00)	Fair
NFI (Normed Fit) Index)	>0.9	0.84	Fair
RMSEA (Root Mean Square Error of Approximation)	<=0.08	0.064	Good
RFI (Relative Fit Index)	>0.90	0.80	Fair
AGFI (Adjusted Goodness of Fit Index)	>0.90	0.77	Fair
CAIC (Consistent Akaike's Information Criterion)	<CAIC for saturated or Independence	CAIC = 814.25 (CAIC saturated=1803.191; CAIC independence =2604.94)	Good
ECVI (Expected Cross-Validation Index)	<ECVI for saturated or Independence	ECVI = 3.45 (ECVI saturated=4.03; ECVI independence =16.84)	Good
TLI / NNFI (Non-Normed Fit Index)	>=0.90	0.91	Good
IFI (Incremental Fit Index)	>90	0.93	Good
CFI (Comparative Fit Index)	>0.9	0.92	Good
GFI (Goodness of Fit Index)	approaching the value of 1 atau >0.9	0.83	Good
SRMR (Standardized RMR)	<0.05	0.081	Fair
PGFI (Parsimony Goodness of Fit Index)	>0.5	0.62	Good
RMR (Root Mean Square Residual)	approaching the value of 0	0.16	Good
PNFI (Parsimony Normed Fit Index)	>0.5	0.69	Good
CMIN / Df	<=3	1.741	Good
CN (Critical N)	>200	106.46	fair

E. Post-Hoc Statistical Power Analysis

Given that all six hypotheses were not supported, a post-hoc statistical power analysis was conducted to assess whether the null results reflect genuine absence of effects or insufficient statistical power. Using G*Power with the parameters of N = 150, $\alpha = 0.05$, and effect sizes derived from comparable prior studies ($r = 0.15$ to $r = 0.25$, corresponding to small-to-medium effects in social science research), the estimated statistical power (β) ranged from approximately 0.18 to 0.42. These values fall substantially below the conventional threshold of 0.80, indicating that the study was underpowered to reliably detect effects of these magnitudes. This means the probability of a Type II error (failing to detect a true effect) was high. Consequently, the null results should be interpreted with caution: they may reflect genuine absence of relationships in this population, or they may be an artifact of insufficient statistical power. A sample of approximately 264–400 respondents would be required to achieve power ≥ 0.80 for the smallest expected effect sizes. Future research should prioritize larger samples to enable definitive conclusions.

VI. DISCUSSION

A. Interpretation of Null Findings

The rejection of all six hypotheses is a notable and theoretically meaningful outcome that warrants careful interpretation. The null results contrast with findings from foundational studies in the HBM-phishing literature [6][10][11][15][16][24]. At least three classes of explanation are plausible. First, the null results may reflect genuine contextual differences: psychological constructs validated in Western or controlled academic settings may not operate the same way among Indonesian general internet users who encounter real-world phishing. Cultural collectivism [3], differential digital literacy [1], and Indonesia's unique phishing landscape (e.g.,

WhatsApp and Instagram-based attacks [2]) may collectively moderate or suppress the relationships hypothesized by the HBM. Second, as demonstrated by the post-hoc power analysis (Section V E), the sample size of 150 was likely insufficient to detect small-to-medium effects, raising the probability of Type II errors across all six hypotheses. Third, two constructs (PBR, PBN) exhibited sub-threshold reliability (Cronbach's $\alpha < 0.60$), which attenuates structural path estimates and reduces the likelihood of detecting significant relationships even where true effects exist.

B. Comparison with Prior Literature

The present findings diverge from several key prior studies. Gwenhure [6] found that perceived severity and self-efficacy significantly predicted security behavior in an HBM model among university students. Adane and Beyene [24] found that perceived barriers significantly influence phishing-related security behavior. Chen et al. [11] established desensitization and past detection success as predictors of susceptibility. The divergence from these findings may be attributed to differences in sample characteristics (students vs. general population), the operationalization of constructs (the current instruments were adapted from English-language scales without full cross-cultural validation), and the measurement of susceptibility (self-report vs. behavioral measures). Notably, the Indonesian population sampled here consists of individuals who had already encountered phishing attacks, potentially introducing selection bias: those who survived prior attacks may have developed heuristic defenses that render the HBM constructs less predictive.

C. The Role of Unmeasured Mediating and Moderating Variables

A critical observation from this study is that the hypothesized direct effects of HBM constructs on phishing susceptibility were uniformly non-significant. This pattern may indicate that the relationships between these constructs and susceptibility are not direct but are mediated or moderated by variables not included in the current model. Digital literacy is a particularly compelling candidate: individuals with higher digital literacy may translate threat perceptions into protective behaviors more effectively [1], creating a suppressor effect on the direct PST–PS path. Cultural norms (e.g., collectivism, deference to authority) may moderate the extent to which perceived benefits or barriers influence behavior [3]. Attack-specific familiarity (that is, knowledge of the specific phishing tactics prevalent in Indonesia) may also moderate the effectiveness of self-efficacy as a predictor. Future studies should incorporate these variables as explicit mediators or moderators within extended HBM or Protection Motivation Theory frameworks.

D. Methodological Limitations and Recommendations

Several methodological limitations of this study should be acknowledged. The sample size of 150 respondents was insufficient for CB-SEM to achieve adequate statistical power for the hypothesized effect sizes, and the CN value of 106.46 confirmed this. The sub-threshold reliability of PBR and PBN indicates that these constructs were not adequately captured by the adapted measurement items in this population. Offline data collection in a single province (Riau) limits the generalizability of findings to the broader Indonesian population. Furthermore, all variables were measured by self-report, introducing the possibility of common method variance. Future research should address these limitations by: 1) expanding the sample to at least 300 respondents drawn from multiple Indonesian provinces; 2) revising or replacing the PBR and PBN items to improve cross-cultural validity; 3) incorporating behavioral or objective phishing susceptibility measures alongside self-report; and 4) including digital literacy, cultural orientation, and attack-type familiarity as additional constructs.

E. Implications

Despite the null hypothesis results, this study offers important practical insights. The finding that none of the measured psychological constructs significantly predict susceptibility suggests that standard awareness-raising interventions targeting threat perception or self-efficacy alone may be insufficient for the Indonesian context. Practitioners (including internet service providers (ISPs), the BSSN, and government agencies) should consider multi-layered interventions that address digital literacy gaps, provide culturally adapted phishing training, and integrate technical defenses with behavioral change programs. The high variability in PS scores (SD = 6.884) indicates that susceptibility is not uniformly distributed, suggesting that targeted interventions for high-risk subpopulations may be more effective than broad awareness campaigns.

VII. Conclusion

This study tested six hypotheses (H1–H6) predicting a positive correlation between phishing susceptibility (PS) and six psychological/perceptual variables, namely perceived severity (PST), perceived barriers (PBR), perceived benefits (PBN), self-efficacy (SE), past success in detection (PSD), and phishing desensitization (PD). Although prior research suggested these relationships, empirical analysis using Structural Equation Modeling (SEM) did not support any of the predicted correlations.

The findings indicate that psychological factors – often assumed to noticeably influence phishing vulnerability – showed no measurable direct effect among the respondents. This suggests that susceptibility may not be driven solely by perceptions but could involve other dimensions, such as digital literacy, social context, cultural security norms, or evolving attack strategies. These possibilities require further investigation rather than being interpreted as confirmed causes. In short, even individuals who perceive phishing threats as severe or have previous success in detection do not necessarily exhibit safer behavior.

The SEM model demonstrated an acceptable overall fit, with several indicators (RMSEA, TLI/NNFI, IFI, and CFI) meeting recommended thresholds, indicating adequate data representation. Parsimony indices (PGFI, PNFI) also confirmed the efficiency of the model, and the low CMIN/df ratio showed a balance between fit and complexity. However, weaknesses in NFI, RFI, AGFI, SRMR, and Critical N suggest the need for refinement – such as increasing sample size, improving the measurement model, and incorporating additional latent variables.

Overall, this study contributes to cybersecurity literature by challenging the assumption that higher threat perception or self-efficacy automatically reduces phishing vulnerability. Future research should investigate mediating or moderating variables that explain the gap between perception and behavior, enrich models with behavioral and contextual factors, and aim for designs that achieve both statistical robustness and real-world relevance in an era of evolving cyber threats.

Future research should expand sample sizes to at least 300 respondents, revise the PBR and PBN measurement items for the Indonesian context, incorporate digital literacy and cultural orientation as mediators or moderators, and consider behavioral phishing susceptibility measures alongside self-report instruments. Theoretically, extending the model to include Protection Motivation Theory constructs or cultural dimensions may improve explanatory power. Overall, this study contributes to the cybersecurity literature by demonstrating that standard HBM-derived constructs do not directly predict phishing susceptibility in the Indonesian general population context, and by providing a foundation for more contextually sensitive models of phishing vulnerability.

REFERENCES

[1] D. B. Utoyo, "Empowering Rural Communities through Digital Literacy: An Empirical Study of Former Women Migrant Workers in Panjer Village, Kebumen - Indonesia," in Proceedings of the 9th International Conference on Indonesian Social and Political Enquiries (ICISPE 2024), vol. 937, in Advances in Social Science, Education and Humanities Research, vol. 937. , Paris: Atlantis Press SARL, 2025, pp. 378–390. doi: 10.2991/978-2-38476-436-5_33.

[2] Rebecca Ling Ze Siew, Brendan Chan Kah Le, Lee Kai Yue, Nuri Nazirah Binti Ismail, Xavier Liong Zhi Hao, and Muhammad Faisal, "Enhancing Security in Industrial IoT: Authentication Solutions Leveraging Blockchain Technology," IJCTS, vol. 1, no. 3, pp. 87–105, Jul. 2024, doi: 10.62951/ijcts.v1i3.29.

[3] A. B. Yunas, Y. Rumanto, and N. Albart, "Blockchain and Green Sukuk Integration for Coastal Community Empowerment in Sustainable Blue Economy," Research Horizon, vol. 5, no. 2, pp. 415–426, 2025.

[4] R. J. Mejias, J. J. Greer, G. C. Greer, M. M. Shepherd, and R. Y. Reyes, "A model for information security vulnerability awareness," Computers & Security, vol. 151, p. 104305, Apr. 2025, doi: 10.1016/j.cose.2024.104305.

[5] R. Rohan, D. Pal, J. Hautamäki, S. Funilkul, W. Chutimaskul, and H. Thapliyal, "A systematic literature review of cybersecurity scales assessing information security awareness," Heliyon, vol. 9, no. 3, p. e14234, Mar. 2023, doi: 10.1016/j.heliyon.2023.e14234.

[6] A. K. Gwenhure, "University students' security behavior against email phishing attacks: insights from the health belief model," Journal of Cybersecurity, vol. 11, no. 1, p. tyaf034, Jan. 2025, doi: 10.1093/cybsec/tyaf034.

[7] K. D. Mitnick and W. L. Simon, The art of deception: Controlling the human element of security. Indianapolis, Indiana: Wiley, 2003.

[8] I. Ghafir et al., "Security threats to critical infrastructure: the human factor," The Journal of Supercomputing, vol. 74, no. 10, pp. 4986–5002, Oct. 2018, doi: 10.1007/s11227-018-2337-2.

[9] S. Anawar, D. L. Kunasegaran, M. Z. Mas'Ud, and N. A. Zakaria, "Analysis of phishing susceptibility in a workplace: A big-five personality perspectives," Journal of Engineering Science and Technology, vol. 14, no. 5, pp. 2865–2882, 2019.

[10] A. Alturki, N. Alshwihi, and A. Algarni, "Factors Influencing Players' Susceptibility to Social Engineering in Social Gaming Networks," IEEE Access, vol. 8, pp. 97383–97391, 2020, doi: 10.1109/ACCESS.2020.2995619.

[11] R. Chen, J. Gaia, and H. R. Rao, "An examination of the effect of recent phishing encounters on phishing susceptibility," Decision Support Systems, vol. 133, p. 113287, Jun. 2020, doi: 10.1016/j.dss.2020.113287.

[12] T. Halevi, J. Lewis, and N. Memon, "A pilot study of cyber security and privacy related behavior and personality traits," in Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro Brazil: ACM, May 2013, pp. 737–744. doi: 10.1145/2487788.2488034.

[13] I. Tjostheim and J. Waterworth, "Exploring susceptibility to phishing: The Cognitive Reflection Test and other possible predictors," in Proceedings of the 57th Hawaii International Conference on System Sciences | 2024, diva-portal.org, 2024. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1823232

[14] E. J. Williams and A. N. Joinson, "Developing a measure of information seeking about phishing," Journal of Cybersecurity, vol. 6, no. 1, p. tyaa001, Jan. 2020, doi: 10.1093/cybsec/tyaa001.

[15] Y. Y. Lee, C. L. Gan, and T. W. Liew, "Thwarting Instant Messaging Phishing Attacks: The Role of Self-Efficacy and the Mediating Effect of Attitude towards Online Sharing of Personal Information," IJERPH, vol. 20, no. 4, p. 3514, Feb. 2023, doi: 10.3390/ijerph20043514.

[16] L. Ribeiro, I. S. Guedes, and C. S. Cardoso, "Which factors predict susceptibility to phishing? An empirical study," Computers &Security. Elsevier, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404823004686

[17] K. Singh, P. Aggarwal, P. Rajivan, and C. Gonzalez, "Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 63, no. 1, pp. 453–457, Nov. 2019, doi: 10.1177/1071181319631355.

[18] S. Baki and R. M. Verma, "Sixteen Years of Phishing User Studies: What Have We Learned?," IEEE Trans. Dependable and Secure Comput., vol. 20, no. 2, pp. 1200–1212, Mar. 2023, doi: 10.1109/TDSC.2022.3151103.

[19] S. Kleitman, M. K. H. Law, and J. Kay, "It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling," PLoS ONE, vol. 13, no. 10, p. e0205089, Oct. 2018, doi: 10.1371/journal.pone.0205089.

[20] L. Q. Ribeiro, I. Guedes, and C. Cardoso, "Phishing: A Theoretical Approach and the Innovative Tools," in Advances in Digital Crime, Forensics, and Cyber Terrorism, N. Mateus-Coelho and M. Cruz-Cunha, Eds., IGI Global, 2023, pp. 76–93. doi: 10.4018/978-1-6684-8422-7.ch005.

[21] A. Sumner, X. Yuan, M. Anwar, and M. McBride, "Examining Factors Impacting the Effectiveness of Anti-Phishing Trainings," Journal of Computer Information Systems, vol. 62, no. 5, pp. 975–997, Sep. 2022, doi: 10.1080/08874417.2021.1955638.

[22] K. S. Jones, N. R. Lodinger, B. P. Widlus, A. S. Namin, E. Maw, and M. Armstrong, "Grouping and Determining Perceived Severity of Cyber-Attack Consequences: Gaining Information Needed to Sonify Cyber-Attacks," J Multimodal User Interfaces, vol. 16, no. 4, pp. 399–412, Dec. 2022, doi: 10.1007/s12193-022-00397-z.

[23] C. L. Anderson and R. Agarwal, "Practicing Safe Computing: A Multimethod Empirical Examination of Home Computer User Security Behavioral Intentions1," MIS Quarterly, vol. 34, no. 3, pp. 613–643, Sep. 2010, doi: 10.2307/25750694.

[24] Kibreab adane and B. Beyene, "Email and Website-Based Phishing Attack: Examining Online Users Security Behavior in Cyberspace Environment," IJISM, vol. 21, no. 1, Jan. 2023, doi: 10.22034/ijism.2022.1977800.0.

[25] R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Research Note —Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance," Information Systems Research, vol. 25, no. 2, pp. 385–400, Jun. 2014, doi: 10.1287/isre.2014.0522.

[26] C. B. Kirwan, D. K. Bjornn, B. B. Anderson, A. Vance, D. Eargle, and J. L. Jenkins, "Repetition of Computer Security Warnings Results in Differential Repetition Suppression Effects as Revealed With Functional MRI," Front. Psychol., vol. 11, p. 528079, Dec. 2020, doi: 10.3389/fpsyg.2020.528079.

AgricultureEarlyWarning: A Web-Based Climate Advisory and Early Warning Platform for Albanian Farmers

Irida Gjermeni
Faculty of Economy and Agribusiness-Department of Mathematics and Informatics,
Agricultural University of Tirana, Tirana, Albania

***Abstract*—This study reports the design and implementation of AgricultureEarlyWarning, a web-based prototype developed to operationalize climate information services and early warning logic for farmer-centered agricultural risk management in Albania. The system integrates parcel-level registration and geolocation, daily and hourly forecast ingestion, crop-stage-sensitive risk evaluation, satellite-derived indicators, dashboard analytics, AI-assisted agronomic explanation, vulnerability profiling, escalation support, and scheduled alert dissemination within a single application. The implementation stack combines ASP.NET MVC on .NET Framework 4.8, SQL Server persistence, Hangfire background scheduling, Open-Meteo weather services, Sentinel Hub APIs, and OpenStreetMap reverse geocoding. Rather than proposing a novel forecasting algorithm or claiming a complete national early warning system, the study contributes an implementation-oriented platform that links data acquisition, contextual risk interpretation, farmer-facing visualization, and asynchronous communication. The findings indicate that the prototype supports localized monitoring, advisory generation, alert scheduling, and analytics, while field validation, long-term adoption assessment, and full institutional response governance remain priorities for future work.**

***Keywords*—AgriculturalEarlyWarning system; climate information services; farmer advisory; remote sensing; ASP.NET MVC; Albania; decision support**

I. Introduction

Climate variability, extreme weather, and crop-specific sensitivity to temperature, rainfall, humidity, and wind continue to challenge agricultural production and farm-level decision-making. In this context, early warning systems (EWSs) are expected to do more than produce forecasts. A modern EWS is an integrated system that combines hazard monitoring and forecasting, risk assessment, communication, and preparedness processes so that people can take timely action before harmful events occur [1]. International frameworks for multi-hazard and people-centered EWSs emphasize four interconnected pillars: risk knowledge, detection and monitoring, warning dissemination and communication, and preparedness to respond [2], [3].

In agriculture, the practical value of climate services depends not only on the scientific quality of the forecast, but also on how well the information is translated into usable decisions. Previous work has shown that climate-service design must be aligned with user needs, decision timescales, and communication channels [4]. In smallholder and mixed farming contexts, mobile delivery channels such as SMS, phone-based interfaces, and localized advisories are particularly important, but their effectiveness still depends on trust, relevance, language, and the capacity of farmers to act on the received information [5]. This means that agricultural EWS design should be farmer-facing, action-oriented, and operationally sustainable.

At the same time, recent literature shows growing interest in agricultural platforms that integrate fine-scale weather information, crop or hazard thresholds, mobile interfaces, digital advisory services, and risk-reduction workflows. Examples include farm-specific agrometeorological warning systems, nationwide agricultural disaster prevention architectures, forecast-driven disease warning systems, and digital advisory initiatives for smallholder agriculture [6]-[9], [19], [20]. These studies reinforce a common lesson: useful agricultural EWSs must connect scientific data, local context, and usable delivery mechanisms.

This study presents AgricultureEarlyWarning, a web-based application developed as a practical prototype for Albanian agriculture. The platform was designed to bridge climate information services (CIS) and agricultural early warning workflows within a single environment. Rather than functioning only as a weather dashboard, the application combines parcel registration, forecast retrieval, crop-aware risk interpretation, climate analytics, AI-assisted agronomic guidance, and scheduled dissemination of daily and event-based alerts. The interface is designed for farmer use, while the architecture also supports administrative oversight and future institutional escalation modules.

The contribution of this study is threefold. First, it documents the architecture and implementation logic of a deployable farmer-oriented agriculture app that integrates multiple open and commercial-ready services. Second, it maps the prototype explicitly to the core pillars of effective EWS design, allowing the reader to assess which components are already operational and which still require institutional completion. Third, it provides a realistic discussion of prototype maturity, including comparison with related agricultural EWSs, prototype-level functional validation, and current limitations. The study is, therefore, framed as a systems design and implementation study, suitable for practitioners and researchers interested in digital agriculture, climate services, and operational EWS development.

The remainder of the study is organized as follows: Section II reviews related work and motivates the design rationale. Section III explains the system requirements and EWS alignment. Section IV presents the architecture. Section V describes the prototype implementation. Section VI

discusses the functional evaluation and practical implications. Section VII outlines limitations and future work, and Section VIII concludes the study.

II. Related Work and Design Rationale

Climate services research consistently shows that information alone is insufficient if it is not embedded in a usable decision context. Born et al. demonstrated through a global meta-analysis that agricultural climate-service use varies by decision type, forecast horizon, and variable, highlighting the importance of matching information products to real farmer decisions [4]. Charalampopoulos and Droulia further stress that climate services for agriculture should move toward integrated, user-oriented pathways that link data generation to sector-specific application needs [6]. These arguments are directly relevant for agricultural apps, where interface simplicity and actionability can determine whether information is adopted or ignored.

Communication design is equally important. Yegbemey and Egah found that mobile-based dissemination is widely used in climate-service delivery to smallholder farmers, but relevance depends on socio-economic fit, literacy, trust in the information source, and the ability of farmers to interpret and respond to messages [5]. This literature supports the use of concise alerts, localized advice, and interfaces in the user's working language. For AgricultureEarlyWarning, this insight motivated the use of Albanian-language dashboards and SMS-oriented dissemination logic.

Recent agricultural EWS implementations also provide useful architectural cues. Shin et al. describe a farm-specific early warning service customized for farmers and extension workers, emphasizing cloud-based coordination, farm-level information delivery, and notification workflows [7]. Yao et al. present a broader agricultural disaster prevention system that integrates forecast services, crop-specific warning logic, and information platforms for farmers and public actors [8]. Krupnik et al. show that a weather-driven wheat blast EWS achieved practical scale only because forecast models, user-centered interfaces, dissemination channels, and institutional cooperation were co-developed [9]. Together, these studies suggest that effective agricultural EWSs require four qualities: localization, interpretability, dissemination, and operational governance.

AgricultureEarlyWarning was designed in response to these lessons and to the needs of a context where farmers may require one integrated entry point for weather, risk, and agronomic interpretation. The design rationale, therefore, prioritizes: 1) parcel-specific localization rather than generic regional weather pages; 2) crop-aware risk interpretation rather than raw meteorological display; 3) actionable advisory text rather than only numerical outputs; 4) support for asynchronous daily and event-based dissemination; and 5) a modular architecture that can evolve from a farmer dashboard into a broader institutional EWS workflow.

Unlike many research prototypes that focus on a single hazard, crop, or model, AgricultureEarlyWarning aims to function as a flexible platform. It combines weather forecasts, selected remote-sensing indicators, crop-stage-sensitive logic, message scheduling, subscription management, and climate analytics in one web system. This broad scope increases implementation complexity, but it also reflects a real operational challenge: farmers do not experience data streams separately. They experience one decision environment in which weather, crop status, risk, timing, and advisory actions need to converge (see Table I).

TABLE I. Comparative Positioning of AgriculturalEarlyWarning and Advisory Systems

System or study	Main focus	Localization	Dissemination	Validation or evidence	Position relative to AgricultureEarlyWarning
Shin et al. [7]	Farm-specific agrometeorological disaster warning service	Farm and extension-worker scale	Service notifications and cloud coordination	Operational service description	Stronger institutional service model; less emphasis on Albanian-language farmer dashboard and subscription workflow
Yao et al. [8]	Agricultural disaster prevention system and adaptation strategy	National/regional platform logic	Information platform for public and farmer actors	System and policy-oriented implementation	Broader governance scope; AgricultureEarlyWarning is lighter and more farmer-facing
Krupnik et al. [9]	Weather-driven wheat blast early warning system	Crop- and disease-specific local warnings	User-centered interfaces and dissemination	User-centered design, validation, and scaling	Stronger disease-specific validation; AgricultureEarlyWarning is multi-risk and modular but less field-validated
AgricultureEarly Warning	Web-based climate advisory and early-warning workflow for Albanian farmers	Parcel coordinates, crop association, vulnerability profile	Dashboard, daily advisory SMS, event-based alerts	Prototype-level functional validation and interface evidence	Combines CIS, parcel registration, risk logic, AI explanation, SMS, analytics, and escalation support in one modular prototype

III. System Requirements and Alignment with Early Warning System Principles

The conceptual starting point of AgricultureEarlyWarning is that an agriculture app should not be evaluated only as a software product, but as a partial operational EWS. From this perspective, the system requirements were derived from the four core EWS pillars [1]-[3].

The first requirement is risk knowledge. The application must represent the local production context in a structured way. For this reason, the system begins with farmer registration and map-based parcel localization. Each farmer is associated with a geographic position, crop information, and cultivation context. This creates the basis for place-specific forecast retrieval and crop-sensitive interpretation. Risk knowledge is also supported by crop parameter definitions, stage-sensitive thresholds, risk

categories, and historical analytics that help transform environmental variables into meaningful agricultural signals.

The second requirement is detection, monitoring, analysis, and forecasting. The platform, therefore, integrates short- to medium-range weather forecast data and selected Earth observation products. Weather variables such as temperature, humidity, rainfall, wind speed, and precipitation probability are retrieved through Open-Meteo APIs [10]. Satellite-related indicators are designed through Sentinel Hub services, where the Processing API can provide imagery and the Statistical API can provide pixel-level summaries such as vegetation and moisture-related indicators without requiring full raster download [11]-[13]. These data sources are combined with rule-based crop logic to derive risk levels and advisory cues.

The third requirement is warning dissemination and communication. The application implements this by coupling dashboards with scheduled and event-driven messaging. Warnings must be timely, understandable, and actionable, not merely visible inside a control panel. This is why the platform supports daily advisory dissemination and repeated background checks for event-based alerts. Message content is designed to summarize the relevant risk, affected crop context, and a practical recommendation. This design is consistent with literature showing the importance of suitable delivery channels and understandable communication for smallholder-oriented climate services [5].

The fourth requirement is preparedness and response capability. This is the area where the current prototype is only partially complete. The farmer-facing side already supports a response-oriented workflow by translating risk scores into suggested actions. However, a full institutional response chain requires escalation management, acknowledgement workflows, case tracking, and formal roles for agronomists or administrators. AgricultureEarlyWarning includes the architectural direction for such modules, but not all of them are fully operational in the current implementation. This distinction is important because many digital platforms overclaim EWS completeness when they mainly cover monitoring and communication. In this study, preparedness is treated as an area of ongoing extension rather than a finished component.

A further requirement concerns sustainability and usability. The system is designed as a web application so that it can operate through common browsers and mobile-responsive interfaces. In addition, it includes a free trial and subscription workflow intended to support a realistic service model for climate information delivery. While economic adoption is outside the scope of the present systems paper, service sustainability is relevant because an EWS that cannot be maintained operationally will not achieve long-term value.

Finally, the system requires trustworthy geospatial localization. Reverse geocoding through Nominatim is used to transform coordinates into readable place references, while respecting API usage requirements and application identification rules [14], [15]. This supports clearer farmer records, contextual dashboards, and location-aware interpretation. Fig. 1 and Table II provide a compact visual interpretation of how these functions relate to the four pillars of an effective early warning system.

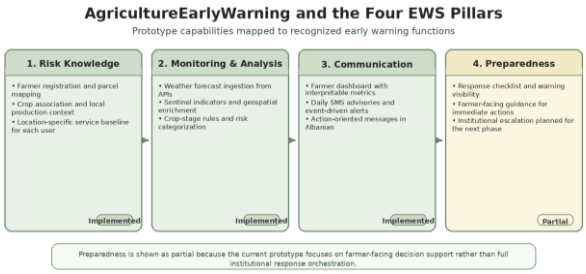


Fig. 1. Alignment of agriculturereearlywarning capabilities with the four pillars of an effective early warning system.

TABLE II. ALIGNMENT OF AGRICULTUREEARLYWARNING WITH EWS PILLARS.

EWS pillar	Current support in prototype	Maturity
Risk knowledge	Parcel registration, crop association, crop parameters, location context, historical analytics	Strong
Detection and monitoring	Forecast ingestion, hourly/daily variables, satellite indicators, threshold logic	Strong
Dissemination and communication	Dashboard, daily advisory messages, event-based alerts, Albanian-language UI	Strong
Preparedness and response	Action suggestions available; institutional escalation and case workflows only partially implemented	Partial

IV. SYSTEM ARCHITECTURE

The overall architecture of AgricultureEarlyWarning is organized into five logical layers: data acquisition, geospatial and farmer context, risk and intelligence processing, application services, and dissemination. Fig. 2 summarizes this structure.

At the data-acquisition layer, the platform retrieves forecast and historical weather variables from Open-Meteo [10]. It is designed to request daily and hourly variables relevant to agriculture, including maximum and minimum temperature, relative humidity, rainfall, precipitation probability, and wind speed. In parallel, satellite-oriented services are integrated through Sentinel Hub APIs [11]-[13]. Sentinel-2 derived vegetation information, such as NDVI, contributes to crop condition interpretation, while Sentinel-1 derived signals such as VV and VH are used as proxies for moisture and crop structural conditions where appropriate. The goal is not to replace agronomic field observations, but to enrich the evidence available for localized advisory logic.

The geospatial and farmer-context layer stores farmer identity, parcel coordinates, crop associations, and administrative location fields. Parcel position is central because all downstream analysis is location-specific. Reverse geocoding adds readable place labels to coordinate-based records [14]. This layer also stores subscription state, trial status, and basic engagement-related attributes needed for service administration.

Above this layer sits the application-services layer, implemented in ASP.NET MVC on .NET Framework 4.8 [18] with SQL Server persistence and background scheduling via Hangfire [16], [17]. This layer exposes the farmer dashboard, registration workflows, climate analytics pages, administrative

controls, subscription handling, escalation support, and AI-assisted chat interface. Background jobs are used for recurrent weather checks and scheduled dissemination. The use of persistent job scheduling is important because warning systems must continue routine monitoring independently of individual user sessions.

The dissemination layer delivers outputs to end users. The primary interface is the responsive web dashboard, which provides current conditions, forecast trajectories, risk summaries, analytics, and recommendations. A second dissemination path uses SMS-style notifications for daily information and event-based alerts. This separation is intentional: a dashboard supports exploration, while alerts support timely action. In operational EWS terms, the dashboard corresponds more to situational awareness, whereas alerts correspond to communication and triggering.

The architecture was deliberately designed to remain modular. Each upstream service can be improved or replaced without rewriting the entire platform. For example, forecast providers can change, satellite workflows can be expanded, and the advisory-generation component can evolve from rule-based logic toward hybrid AI-assisted interpretation. This modularity is essential for deployment in real settings, where data availability, cost, and infrastructure constraints may change over time. The logical relationships between layers are visualized in Fig. 2.

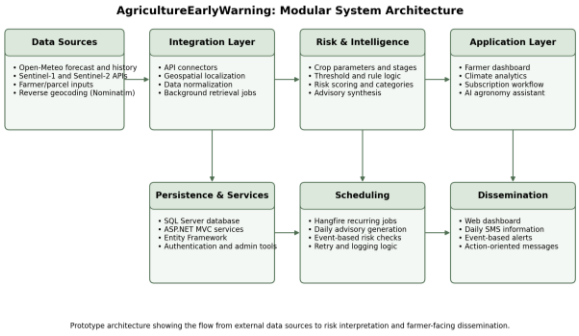


Fig. 2. Modular architecture of the AgricultureEarlyWarning prototype.

V. Prototype Implementation and Interface Realization

The current implementation of AgricultureEarlyWarning is realized as an ASP.NET MVC application on .NET Framework 4.8 with SQL Server persistence and Hangfire-based background processing. This section documents the main functional modules and provides interface-level evidence of how the prototype operationalizes registration, monitoring, analytics, communication, and preparedness-support functions. To improve technical transparency, the implementation is described at four levels: database structure, risk scoring logic, alert rules, and API integration (see Table III).

At the database level, the prototype uses normalized SQL Server entities to connect users, parcels, crops, warning records, SMS logs, vulnerability profiles, escalation events, and feedback. The database design separates static registration data from time-varying climate-risk records so that monitoring results can be audited and reused in analytics.

TABLE III. Technical Implementation Details of the Prototype

Layer	Main implementation elements	Operational role
Database structure	Farmers, Crops, FarmerCrops, ClimateRiskPredictions, Sms Alerts, FarmerVulnerability Profiles, EwsAlertRecords, EwsEscalationEvents, Ews DispatchLogs, EwsFeedbacks, and EwsInstitutionalCases	Stores parcel context, crop associations, computed warnings, communication history, vulnerability data, escalation status, and feedback.
Risk scoring logic	Rule-based thresholds for temperature, rainfall, humidity, wind, crop stage, vulnerability score, and advisory category	Converts forecast variables into a 0-100 risk score and qualitative warning levels such as green, yellow, orange, and red.
Alert rules	Daily advisory job, event-based checks, acknowledgement status, escalation delay, and retry/logging logic	Supports asynchronous warning dissemination and prepares the system for institutional escalation workflows.
API integration	Open-Meteo forecast requests, Sentinel Hub Processing/Statistical APIs, Nominatim reverse geocoding, and Hangfire scheduling services	Combines external data acquisition with geospatial localization, remote-sensing enrichment, and recurring background execution.
AI explanation layer	Server-side AIChatService with constrained prompt template and deterministic risk facts supplied by the warning engine	Explains warnings and general management actions without changing the computed risk level or acting as the trigger mechanism.

The risk score is calculated as an interpretable engineering score rather than as a learned statistical forecast. Each relevant meteorological variable is normalized against crop- and stage-specific thresholds. The resulting partial scores are multiplied by crop-stage weights and combined with vulnerability modifiers. The final score is mapped into warning levels so that the farmer sees a simple risk category while the administrator can still inspect the underlying variables.

Alert rules are implemented as scheduled and event-triggered routines. The daily advisory routine summarizes expected conditions for the active parcel and crop, while event-based checks are used when current or forecast conditions exceed configured thresholds. The escalation rule screen defines acknowledgement windows and transfer points from farmer notification to agronomist or administrator review. In the current prototype, these workflows support preparedness logic, but formal institutional operating procedures are still required before the platform can be described as a complete institutional EWS.

API integration is isolated behind service classes so that external providers can be changed without rewriting the dashboard or alert modules. Open-Meteo supplies daily and hourly forecast variables [10]. Sentinel Hub services support satellite-derived indicators and statistics [11]-[13]. Nominatim is used for reverse geocoding and readable location labels [14], [15]. Hangfire executes recurring checks and scheduled dissemination independently of user login sessions [16], [17].

A. Farmer Registration and Localization

The onboarding flow allows a farmer to register and select a parcel on the map. Geographic coordinates are stored

directly and then enriched with reverse-geocoded location information. This step is foundational because subsequent monitoring, forecast requests, and risk interpretation are all tied to the selected parcel. The registration process is also connected to free-trial and subscription logic, reflecting the practical requirement that operational climate information services need a sustainable delivery model.

B. Weather and Advisory Dashboard

After registration, the farmer accesses a dashboard that combines forecast information, risk summaries, and practical guidance. The platform supports daily and hourly weather views and presents metrics in agricultural terms rather than as raw meteorological lists. This includes explicit units, intuitive risk badges, and concise explanatory text. The design seeks to reduce interpretation burden for the user while preserving transparency about the underlying variables.

C. Crop-Aware Risk Logic

The risk engine evaluates conditions such as temperature extremes, excessive humidity, rainfall accumulation, and wind intensity relative to crop-stage sensitivity and configured crop thresholds. This allows the same forecast signal to be interpreted differently for different crop and stage combinations. For instance, a humidity and rainfall episode may be treated as a stronger disease risk during sensitive crop stages than during less vulnerable periods. The platform, therefore, behaves as a decision-support layer on top of the forecast rather than as a forecast viewer only.

D. Remote-Sensing Enrichment

The architecture supports the use of Sentinel-derived indicators to complement weather-based interpretation. NDVI can be used as a proxy for vegetation vigor, while Sentinel-1 backscatter-derived values can contribute to moisture and structural interpretation. In the current prototype, these indicators are integrated as additional evidence streams that can enrich dashboards and analytics. Their role is supportive rather than fully autonomous, since agronomic interpretation still benefits from local calibration and field knowledge.

E. AI-Assisted Agronomic Guidance

The system contains an AI-assisted chat component implemented as a server-side AIChatService connected to a configurable large-language-model endpoint. The component is not used to calculate risk scores or trigger alerts. Instead, it receives structured facts from the deterministic warning engine, including crop, stage, location context, relevant weather variables, computed risk level, and the active advisory category. The prompt template constrains the model to explain why the warning appears, translate technical variables into farmer-friendly Albanian, and provide conservative general actions. It also instructs the model not to invent sensor values, not to change the computed risk level, and not to provide pesticide doses unless such advice is grounded in an official label or validated agronomic source. This design follows the emerging view that generative AI can support agricultural advisory communication, but that guardrails, domain facts, and human or expert oversight are necessary for safe deployment [21], [22]. Generated text is, therefore, treated as an interpretability layer, while the operational warning logic remains deterministic and auditable.

F. Background Scheduling and Notification

Routine monitoring and dissemination are handled through Hangfire recurring jobs [16], [17]. This allows the application to execute regular checks, such as periodic weather retrieval and event-driven warning evaluation, as well as scheduled daily information delivery. From an EWS perspective, this is a critical operational feature because alerts should not depend on a farmer being logged into the platform at the moment a hazardous condition develops.

G. Climate Analytics and Administration

Beyond immediate alerts, the prototype supports climate analytics views that aggregate historical records, summary metrics, and time series. This supports reflection and learning, not just immediate action. Administrative features allow oversight of farmer records, crop associations, and other service-management functions. In future iterations, these administrative capabilities are expected to evolve into institutional response and case-management modules.

Fig. 3 synthesizes the implemented operational chain from onboarding and data retrieval to advisory delivery and learning-oriented analytics. Table IV summarizes the main implemented modules, while Fig. 4(a) to Fig. 4(d) and Fig. 5(a) to Fig. 5(d) provide enlarged interface-level evidence from the deployed prototype.

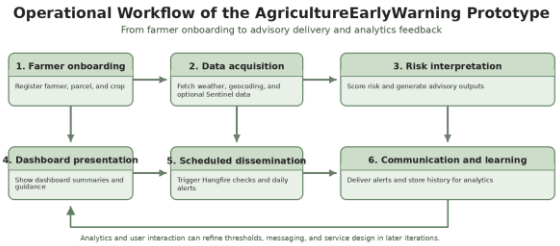


Fig. 3. Farmer-facing operational workflow of the AgricultureEarlyWarning prototype from onboarding to advisory dissemination and analytics feedback.

TABLE IV. MAIN IMPLEMENTED MODULES OF THE PROTOTYPE

Module	Implemented function	Operational note
Farmer onboarding	Registration, parcel geolocation, crop selection, subscription/trial linkage	Creates parcel-specific service context
Weather dashboard	Current, hourly, and daily forecast display with agricultural units and risk badges	Farmer-facing situational awareness
Risk engine	Crop-stage-sensitive thresholds for temperature, humidity, rainfall, and wind	Transforms forecast values into warnings
Remote-sensing enrichment	Sentinel-1 / Sentinel-2 indicator support including NDVI and backscatter summaries	Complements weather-based interpretation
AI guidance	Conversational explanation layer in Albanian	Improves interpretability, not core warning logic
Dissemination	Daily advisory SMS and event-based alerts via recurring jobs	Supports asynchronous communication
Climate analytics	Historical summaries, trends, and export-oriented views	Supports review and learning

H. Interface-Level Implementation Evidence

Fig. 4(a) to Fig. 4(d) and Fig. 5(a) to Fig. 5(d) present enlarged screenshots from the realized prototype. They document how the implementation translates the architectural workflow into concrete user interfaces for parcel geolocation, dashboard monitoring, analytical interpretation, vulnerability assessment, survey capture, escalation management, and farmer communication through SMS-style dissemination.

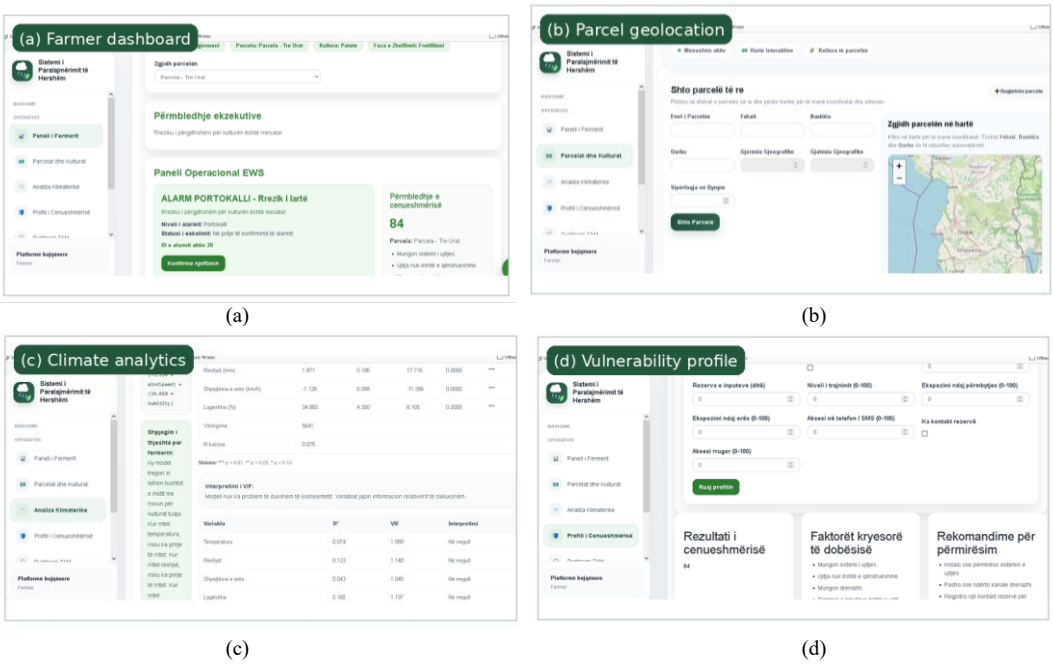


Fig. 4. (a) Farmer dashboard showing parcel selection, current risk summary, acknowledgement status, and vulnerability score. (b) Parcel geolocation screen where the farmer enters parcel attributes and selects coordinates on the interactive map. (c) Climate analytics screen with regression-style summary indicators, VIF interpretation, and climate variable diagnostics. (d) Vulnerability profile screen recording, exposure, access, training, backup capacity, and practical recommendations.

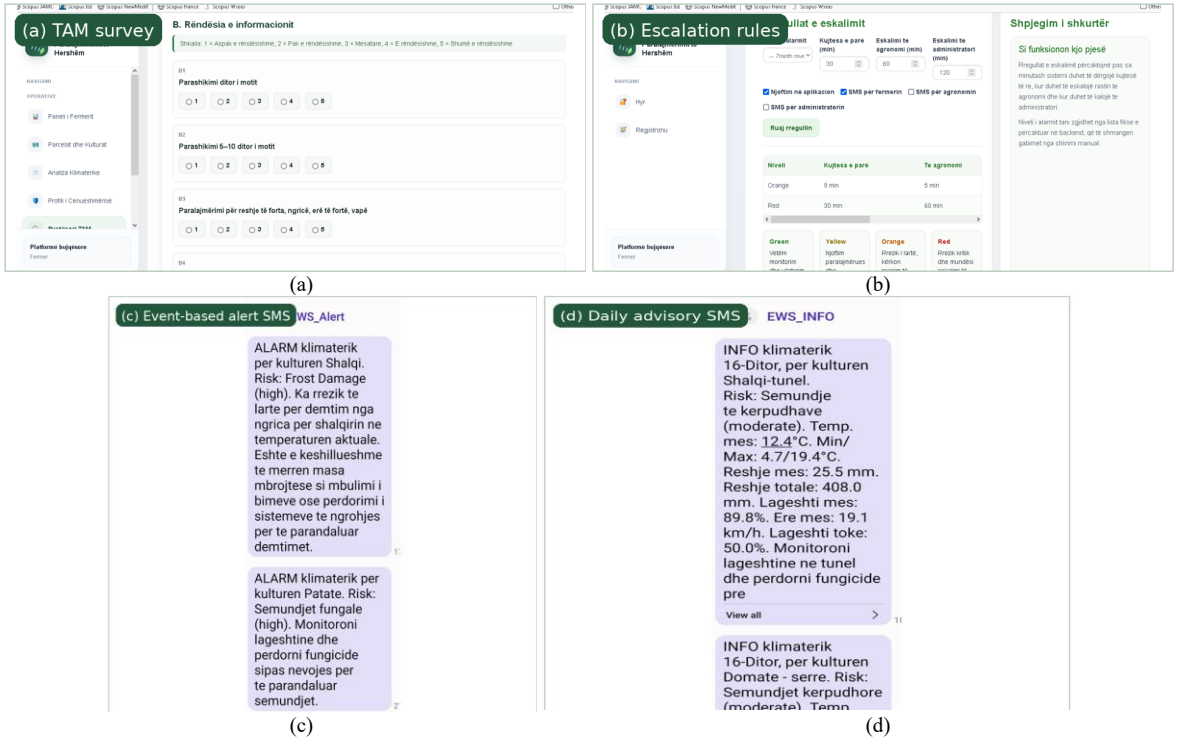


Fig. 5. (a) TAM survey capture screen used to collect structured user perceptions about the advisory service. (b) Escalation-rule configuration screen defining acknowledgement delays, transfer to agronomist, and transfer to administrator. (c) Event-based alert SMS showing risk type, severity, crop context, and immediate recommended action. (d) Daily advisory SMS showing forecast summary, disease-related risk, wind, rainfall, humidity, and suggested monitoring action.

VI. Functional Evaluation and Technical Discussion

This study does not report a randomized field trial or a statistically validated impact assessment. Instead, the revised evaluation reports prototype-level functional validation, architectural coherence, and implementation evidence drawn from the realized workflow and user interfaces. This evaluation mode is appropriate for a design and implementation paper, but it is clearly separated from field effectiveness, adoption, or yield-impact claims.

The functional validation focuses on measurable implementation indicators rather than agronomic impact. Table V summarizes the verification evidence available at prototype stage. The table deliberately distinguishes implemented functionality from performance metrics that require future field deployment, such as response latency under production load, API uptime over a full season, and farmer acceptance after repeated use.

TABLE V. Prototype-Level Functional Validation Indicators

Validation dimension	Measured or verifiable indicator	Prototype evidence	Remaining gap
EWS pillar coverage	Four EWS pillars reviewed; three assessed as strong and one as partial	Table II maps risk knowledge, monitoring, dissemination, and preparedness	Preparedness requires formal institutional operating procedures.
Functional module coverage	Seven implemented module groups plus escalation, feedback, and vulnerability-support screens	Table IV and Fig. 4(a)-Fig. 5(d) document the workflow	Independent usability testing is still needed.
Scheduler configuration	Recurring event checks and daily advisory dissemination configured through Hangfire	Daily EWS_INFO and event-based EWS_ALERT examples are captured in Fig. 5c and Fig. 5d	Long-term scheduler accuracy and latency must be measured under production load.
Risk-output transparency	Risk level, crop context, weather variables, vulnerability score, and advisory text stored or displayed	Dashboard, analytics, SMS, and vulnerability screens expose the key decision variables	Threshold calibration across seasons and crops remains future work.
API integration coverage	Weather, satellite, reverse geocoding, persistence, and background-job services integrated	Open-Meteo, Sentinel Hub, Nominatim, SQL Server, and Hangfire are separated through service classes	API reliability and rate-limit behavior require season-long monitoring.
Communication outputs	Two communication formats: daily advisory and event-based alert	SMS-style messages include crop, risk type, severity, and action-oriented text	Message comprehension and trust should be tested with farmers and extension staff.

The first strength of the prototype is localization. By structuring the workflow around parcel coordinates and crop associations, the system avoids the genericity of broad regional weather dashboards. This is consistent with recent agricultural EWS examples that emphasize farm-specific or location-specific service logic [7]-[9]. The use of remote-sensing services further extends the potential for localized interpretation, especially in settings where in-field instrumentation is limited.

The second strength is dissemination design. The system recognizes that warnings must reach users through appropriate channels and at suitable times. This reflects prior evidence that communication channels are central to climate-service usability for farmers [5]. By combining an exploratory dashboard with daily and event-based alerts, the platform balances information richness with timely communication. The Albanian-language interface also supports contextual usability.

The prototype also has several limitations. First, the agronomic quality of risk outputs depends on the calibration of crop parameters, thresholds, and advisory texts. Without extended field validation across crops, seasons, and agroecological zones, warning quality should be interpreted cautiously. Second, the integration of satellite indicators is promising but still requires rigorous calibration and validation for operational agricultural inference. Third, the platform currently emphasizes the farmer-facing and analytical sides of the EWS more than the institutional response side. It is therefore presented as a farmer-centered prototype with partial preparedness support, not as a complete national or institutional EWS.

The fourth limitation concerns operational infrastructure. Because the system depends on external APIs, background processing, and hosting reliability, production deployment requires attention to service continuity, usage policies, rate limits, scheduler robustness, and data-protection rules [10]-[17]. This is not a conceptual weakness of the design, but it is a real implementation constraint that any operational EWS must manage.

The generalizability of the architecture is strongest where three assumptions hold: geospatial parcel coordinates can be collected, weather and satellite APIs are accessible, and an extension or administrative actor can respond to escalated warnings. The software pattern is transferable because the data-acquisition, risk-logic, dashboard, and dissemination layers are modular. However, thresholds, language, crops, institutions, SMS providers, and payment models must be localized before deployment in another region or country. The platform is therefore best understood as a reusable architectural blueprint rather than a directly universal solution.

VII. Limitations and Future Work

Several development directions emerge from the current prototype and from the limits of the functional validation reported above.

First, field validation should become the priority. Future work should compare risk outputs and advisories with observed crop conditions, farmer actions, and actual hazard

outcomes across seasons. This would enable calibration of thresholds and the estimation of practical utility.

Second, the preparedness and response pillar should be expanded from prototype support into a formal institutional workflow. The current interface includes vulnerability profiling, escalation-rule configuration, feedback capture, and communication examples. Future deployment should add formal alert acknowledgement, agronomist assignment, dispatch logs, case closure rules, and institutional accountability procedures. Such additions would move the system from a strong farmer information platform toward a more complete end-to-end EWS.

Third, remote-sensing integration should be strengthened. Future work can evaluate how NDVI, VV, VH, and related indices improve risk prediction, especially for stress detection, excess moisture interpretation, and disease-supportive conditions. At present, these indicators are best understood as complementary evidence streams.

Fourth, the system would benefit from broader communication options and user testing. SMS remains important, but future versions may include WhatsApp-like channels, richer multilingual content, structured advisory templates, and usability testing with farmers and extension personnel.

Fifth, longer-term economic and adoption evaluation is needed. Since climate information services are also service products, future studies should assess farmer acceptance, repeated use, willingness to pay, and the institutional sustainability of subscription-based or mixed delivery models.

Finally, future technical work may include migration toward newer .NET stacks, stronger API abstraction, automatic performance logging, dashboard latency monitoring, improved audit trails, and richer analytics for both farmers and administrators. These enhancements would improve maintainability and long-term deployment readiness.

VIII. Conclusion

This study presented AgricultureEarlyWarning as a design and implementation case of a web-based climate advisory and early warning platform for farmer-centered agricultural risk management in Albania. The prototype integrates parcel-based localization, weather and satellite data services, crop-aware risk interpretation, dashboard analytics, AI-assisted explanation, vulnerability profiling, escalation support, and scheduled dissemination inside a single operational environment. Its main contribution is the engineering of a modular workflow that links risk knowledge, monitoring, communication, and response support in a form that is implementable with existing web technologies and external data services.

The study shows that a practical agricultural EWS pathway does not need to start as a monolithic national infrastructure. It can begin as a modular, farmer-facing platform that incrementally connects risk knowledge, monitoring, warning communication, and response support. AgricultureEarlyWarning already demonstrates meaningful progress in the first three of these areas and provides prototype-level support for the fourth. Institutional preparedness and coordinated response remain the next major step.

For digital agriculture researchers and practitioners, the platform illustrates how climate information services can be transformed into a more actionable service layer for farming decisions. For implementation contexts similar to Albania, it provides a realistic blueprint for integrating open weather services, satellite APIs, geospatial localization, and warning dissemination into one coherent system. The prototype, therefore, represents a substantive step toward practical, localized, and scalable agricultural early warning support.

References

[1] UNDRR, "Early warning system," UNDRR Terminology, 2009. [Online]. Available: https://www.undrr.org/terminology/early-warning-system. Accessed: Apr. 17, 2026.

[2] World Meteorological Organization, Multi-hazard Early Warning Systems: A Checklist. Geneva, Switzerland: WMO, 2018.

[3] UNDRR, "Early Warnings for All," 2024. [Online]. Available: https://www.undrr.org/implementing-sendai-framework/sendai-framework-action/early-warnings-for-all. Accessed: Apr. 17, 2026.

[4] L. Born, S. Prager, J. Ramirez-Villegas, and P. Imbach, "A global meta-analysis of climate services and decision-making in agriculture," Climate Services, vol. 22, p. 100231, 2021, doi: 10.1016/j.cliser.2021.100231.

[5] R. N. Yegbemey and J. Egah, "Reaching out to smallholder farmers in developing countries with climate services: A literature review of current information delivery channels," Climate Services, vol. 23, p. 100253, 2021, doi: 10.1016/j.cliser.2021.100253.

[6] I. Charalampopoulos and F. Droulia, "A pathway towards climate services for the agricultural sector," Climate, vol. 12, no. 2, p. 18, 2024, doi: 10.3390/cli12020018.

[7] Y.-S. Shin, H.-A. Lee, S.-H. Park, Y.-K. Han, K.-M. Shim, and S.-J. Han, "Establishment and operation of an early warning service for agrometeorological disasters customized for farmers and extension workers at metropolitan-scale," Atmosphere, vol. 16, no. 3, p. 291, 2025, doi: 10.3390/atmos16030291.

[8] M.-H. Yao, Y.-H. Hsu, T.-Y. Li, Y.-M. Chen, C.-T. Lu, C.-L. Chen, and P.-Y. Shih, "Agricultural disaster prevention system: Insights from Taiwan's adaptation strategies," Atmosphere, vol. 15, no. 5, p. 526, 2024, doi: 10.3390/atmos15050526.

[9] [9] T. J. Krupnik et al., "A weather-forecast driven early warning system for wheat blast disease: User-centered design, validation, and scaling in Bangladesh and Brazil," Climate Services, vol. 39, p. 100589, 2025, doi: 10.1016/j.cliser.2025.100589.

[10] P. Zippenfenig, "Open-Meteo.com Weather API," Zenodo, 2023, doi: 10.5281/zenodo.7970649.

[11] Sentinel Hub, "Sentinel Hub API documentation," Sentinel Hub Documentation, Sinergise Ltd., 2026. [Online]. Available: https://docs.sentinel-hub.com/api/latest/. Accessed: Apr. 17, 2026.

[12] Sentinel Hub, "Processing API," Sentinel Hub Documentation, Sinergise Ltd., 2026. [Online]. Available: https://docs.sentinel-hub.com/api/latest/api/process/. Accessed: Apr. 17, 2026.

[13] Sentinel Hub, "Statistical API," Sentinel Hub Documentation, Sinergise Ltd., 2026. [Online]. Available: https://docs.sentinel-hub.com/api/latest/api/statistical/. Accessed: Apr. 17, 2026.

[14] Nominatim, "Reverse geocoding," Nominatim 5.3.1 Manual, OpenStreetMap Foundation, 2026. [Online]. Available: https://nominatim.org/release-docs/latest/api/Reverse/. Accessed: Apr. 17, 2026.

[15] OpenStreetMap Foundation, "Nominatim Usage Policy," OpenStreetMap Operations, 2026. [Online]. Available: https://operations.osmfoundation.org/policies/nominatim/. Accessed: Apr. 17, 2026.

[16] Hangfire, "Performing recurrent tasks," Hangfire Documentation, 2026. [Online]. Available: https://docs.hangfire.io/en/latest/background-methods/performing-recurrent-tasks.html. Accessed: Apr. 17, 2026.

[17] Hangfire, "Making ASP.NET applications always running," Hangfire Documentation, 2026. [Online]. Available: https://docs.hangfire.io/en/latest/deployment-to-production/making-aspnet-app-always-running.html. Accessed: Apr. 17, 2026.

[18] Microsoft, "ASP.NET MVC 5," Microsoft Learn, 2024. [Online]. Available: https://learn.microsoft.com/en-us/aspnet/mvc/overview/getting-started/introduction/getting-started. Accessed: Apr. 17, 2026.

[19] Food and Agriculture Organization of the United Nations, The Impact of Disasters on Agriculture and Food Security 2025: Digital Solutions for Reducing Risks and Impacts. Rome, Italy: FAO, 2025.

[20] AIM for Scale, Digital Advisory Services for Agriculture: Innovation Package. Washington, DC: Agricultural Innovation Mechanism for Scale, 2025.

[21] AIEP Initiative, S. Collis, F. Kinyua, V. Kumar, H. Lakougna, C. Merz, K. Pandey, and C. Resch, Disrupting Agricultural Advisory Utilizing Generative AI: Lessons Learned from Minimum Viable Product Development of the Agricultural Information Exchange Platform. Gates Foundation, GIZ, and CLEAR Global, 2025.

[22] S. Yang, Z. Yuan, S. Li, R. Peng, K. Liu, and P. Yang, "GPT-4 as evaluator: Evaluating large language models on pest management in agriculture," arXiv preprint arXiv:2403.11858, 2024.

Spatial Anemia and Osmotic Fragility of Erythrocytes in Microgravity Conditions: A Systematic Review

Lilian Ocares-Cunyarachi[1], Natalia Vargas-Cuentas[2], Avid Roman-Gonzales[3], Georgina Chávez[4], Brigitte Torres-Pinedo[5], Ana Huamani-Huaracca[6]*
Image Processing Research Laboratory (INTI-Lab), Universidad de Ciencias y Humanidades, Lima, Perú[1, 2, 5, 6]
Aerospace Sciences & Health Research Laboratory (INCAS-Lab), Universidad Nacional Tecnológica de Lima Sur[3]
Coordinación de Investigación, Sede La Paz, Universidad Católica Boliviana "San Pablo"[4]

***Abstract*—The study of the effects of microgravity on the human body has become increasingly relevant in biomedical research, particularly with regard to hematological changes. Therefore, the objective was to analyze the available scientific evidence on the impact of microgravity, both real and simulated, on erythrocytes with emphasis on osmotic fragility, cell membrane stability, and hemolysis. Likewise, 40 articles were selected using the PRISMA methodology, using the Scopus, PubMed, Web of Science, and IEEE Xplore databases, applying inclusion and exclusion criteria that had been previously established. Tools such as VOSviewer were also used, which facilitated the visualization of patterns of collaboration and scientific production. The results showed that microgravity produces remarkable changes in erythrocytes, such as an increase in hemolysis, alterations in the capacity for cell deformation, increased oxidative stress, and variations in the membrane. However, most research shows indirect information when it comes to osmotic fragility, with few applications of standardized measurements such as OFT or MCF-H50. In conclusion, severity influences the integrity and functionality of erythrocytes, leading to spatial anemia, although studies with greater standardization in their methodology are needed.**

***Keywords*—Erythrocytes; hemolysis; microgravity; osmotic fragility; spatial anemia**

I. Introduction

According to a study reported by NASA, technical reports served in 2024 indicated the effects of discharge on hematopoiesis and the subsequent life cycle of red blood cells evaluated at the resolution of the individual cell phone, where it indicates that prolonged exposure to microgravity causes the so-called spatial anemia. Because this phenomenon is associated with the absence of mechanical load, which alters erythropoiesis and can shorten the lifespan of erythrocytes, evidencing the key role of gravity in the regulation of the hematological system [1].

During the last decades, space exploration has been increasing, which has promoted the study of the effects of microgravity on the basis of the human body, in this way generating new challenges in the field of medicine. In this sense, according to the authors, they emphasized that one of the most relevant areas is the analysis of blood samples, due to its importance regarding the evaluation of physiological state; therefore, health monitoring and the diagnosis of diseases arise in spatial and terrestrial conditions [2]. In this way, microgravity produces significant alterations in the human body, such as the distribution of fluids, changes in plasma volume, and modifications in the production and function of blood cells.

Thus, in this study, they emphasized that these variations can directly affect hematological and biochemical parameters, which can make it difficult to interpret the results obtained through conventional methods of clinical analysis. Specifically, erythrocytes play an important role in oxygen transport, and their stability depends on the integrity of their cell membrane. In this regard, osmotic fragility is presented as an essential indicator to evaluate such integrity, since it shows the ability of red blood cells to resist changes in the osmotic medium without undergoing hemolysis [3]. Therefore, different studies have indicated that microgravity conditions can alter these properties, increasing the sensitivity of erythrocytes to destruction.

In this way, due to the limitations to carry out studies in space, tools have been developed, such as the clinostat, allowing to simulate microgravity conditions in the laboratory through the constant rotation of samples, facilitating experimental research in controlled environments, analyzing and identifying the effects of microgravity on human cells without requiring an off-planet study [4]. Likewise, the use of simulated microgravity models has allowed to expand knowledge based on the cellular response with respect to the alteration of the gravitational vector; however, there are still some evidence-based limitations in relation to specific hematological parameters [5].

In order to be able to carry out studies in microgravity, simulated microgravity models have been implemented, for example, the clinostat, which allows similar conditions to be developed in the laboratory by continuously rotating the samples. In this way, they help research into the effects of microgravity in a controlled way and have also contributed to the development of their own studies and scientific training [6]. Several studies have shown that microgravity produces important changes in the body's homeostasis, including alterations in fluid distribution, lower plasma volume, and alterations in blood cell production or function. Within this context, spatial anemia has been made known as one of the physiological strategies adopted in the course of a prolonged space mission. Spatial anemia affects the growth of the erythrocyte mass, maintaining an increase in the destruction of red blood cells [7]. This shows the complex nature of human adaptation to the space environment.

In this context, erythrocytes are essential for oxygen transport and their stability depends on the integrity of their cell

*Corresponding author

membrane. Osmotic fragility appears as a possible marker to assess such integrity, since it evaluates the ability of red blood cells to resist changes in the osmotic environment without undergoing hemolysis. Recently, there is data that points to how the structural and, if not functional, properties of erythrocytes can change under microgravity conditions, which can increase the risk of destruction and contribute to the picture of spatial anemia [8]. Therefore, the importance of the present study is to understand how simulated microgravity conditions can affect the stability of erythrocytes, which is why it is significant both for aerospace medicine and for biomedical research. In this way, it allows the use of methodologies accessible in the laboratory to be enhanced, contributing to the generation of knowledge applicable in contexts where it is possible to access real conditions of microgravity.

Despite the advances made, scientific information on the effects of microgravity on hematological indicators, such as osmotic fragility of erythrocytes, remains limited and scattered in the literature. In particular, there is no systematic synthesis that integrates the available findings on this phenomenon, which hinders a clear and comprehensive understanding of its implications on human health. Based on this problem, the research questions that guided this study are presented: What is the available scientific evidence on the effects of microgravity, real or simulated, on the osmotic fragility of erythrocytes? What methodologies have been used for its study? What are the main limitations and gaps in the current literature?

Therefore, the objective of this systematic review is to analyze and synthesize the available scientific evidence on the effects of microgravity on the osmotic fragility of erythrocytes, identifying the main findings, methodologies used, and limitations existing in the present studies. This study seeks to contribute to a better understanding of the hematological changes associated with microgravity, as well as to provide a solid theoretical basis to guide future research in the field of aerospace medicine and biomedical research.

The study is structured as follows: Section II describes in detail the literature review; Section III presents the materials and methods; then Section IV presents the results, Section V presents the discussion; Section VI presents the conclusion, and finally Section VII presents the recommendations and future work.

II. Literature Review

A. Background

Sickle cell anemia is a genetic disease whose clinical expression varies considerably between individuals, which has been attributed to the interaction between genetic and environmental factors. Recent literature highlights the importance of these factors in modulating disease severity, as well as in response to treatment. In this context, microgravity has been studied as a relevant experimental model, since it alters fundamental cellular processes. Various studies have shown that in these conditions, hemolysis and oxidative stress in erythrocytes are increased, key mechanisms in the pathophysiology of the disease. Likewise, changes in the structure and functionality of red blood cells have been observed, which could aggravate their characteristic alterations. Taken together, these findings suggest that the study of extreme environments such as microgravity contributes to a better understanding of sickle cell anemia and offers new perspectives for the development of more effective therapeutic strategies [9].

The study presented by Agramont et al. [10] argues that osmotic fragility in erythrocytes is related to spatial anemia; that is, they evaluated in detail the behavior of erythrocytes in the face of variations in the osmotic environment, in such a way, it was evidenced that they can experience relevant alterations in their functionality and structure. In this way, the results showed that it is possible to be under conditions that manage to simulate extreme environments such as microgravity. Consequently, erythrocytes are more susceptible to hemolysis, which has an essential impact on cell stability, as well as on the accuracy of hematological analyses. The study was able to highlight these modifications, as they can influence the interpretation of laboratory results, that is, directly in contexts where physical conditions differ from those on land. Therefore, the research provided significant evidence on the need to consider the effects of the environment in the analysis of blood samples, thus contributing to a better understanding of the physiological changes associated with microgravity.

Therefore, a study by Manis et al. [11] emphasizes that the behavior of human erythrocytes under simulated microgravity conditions showed that these cells undergo structural changes. Thus, according to the research, a decrease in antioxidant capacity, increased oxidative stress and alterations in the lipid composition of the cell membrane were obtained as a result. Therefore, modifications in the shape and stability of erythrocytes were identified, suggesting that microgravity can directly affect the integrity of the cell membrane. Thus, these results are relevant, since the stability of the cell membrane is related to the osmotic fragility of erythrocytes.

Thus, Murali and Sarkar [12] studied the behavior of erythrocytes under different gravity conditions, demonstrating that the variation of the gravitational field influences cell shape, deformability and dynamics. The results showed that microgravity generates changes in the structure of erythrocytes, affecting their stability and interaction with the environment, which could increase their susceptibility to hemolysis processes.

However, according to a study by Manis et al. [13], they analyzed human erythrocytes under simulated microgravity conditions, evidencing significant alterations in the lipid composition of the cell membrane. The results showed the presence of oxidized lipids and changes in phospholipids, indicating cellular stress and structural modifications in erythrocytes. These changes directly affect membrane stability, which is related to a possible increase in osmotic fragility.

In this sense, the authors emphasize that hemolysis is a determining factor in the development of anemia during long-duration spaceflight. The authors point out that, under conditions of microgravity, there is a sustained increase in the destruction of red blood cells, which shows that anemia is not only a temporary adaptation, but an active physiological process. The authors also highlight that the high presence of by-products of hemoglobin degradation confirms the persistence of hemolysis during the space mission. They also stress that some of these effects continue even after return to Earth, suggesting a

prolonged impact of the space environment on blood physiology. In this regard, the study concludes that environmental factors, such as microgravity, play a key role in the regulation and survival of red blood cells [14].

In this sense, the authors point out that erythrocytes are not a uniform population, but rather present a remarkable heterogeneity in their structure, function and biochemical properties. The authors point out that these differences may be caused by factors such as cell age, physiological environment and metabolic conditions. The authors also highlight that this variability directly influences the ability of red blood cells to carry oxygen and respond to stressful situations. In this sense, they stress that some erythrocytes may be more susceptible to processes such as hemolysis, which has important implications in the development of various hematological diseases. Similarly, they emphasize that erythrocyte heterogeneity is not only a normal biological phenomenon but can also contribute to the progression of pathologies, affecting the stability and functionality of the blood system [15].

However, in this study, they emphasized exposure to microgravity since it generates multiple physiological alterations at the systemic level, in such a way that the hematological system being one of the most affected. Likewise, they mention that during space missions, a progressive reduction in the mass of erythrocytes was observed, which directly contributes to the development of space anemia. In this way, this phenomenon does not manifest itself in isolation, but is accompanied by changes in the distribution of body fluids, which can mainly mask the real decrease in red blood cells. They also indicate that these changes are linked to the body's adaptive mechanisms in a space environment, in addition to producing a decrease in erythropoietin production and an increase in erythrocyte destruction. It is highlighted that microgravity alters oxygen homeostasis, which forces the body to modify its transport capacity [16].

According to the authors, erythrocytes undergo significant metabolic alterations during spaceflight, demonstrating that microgravity has a direct impact on their biochemical functioning. Thus, changes in essential metabolic pathways, such as energy metabolism, glycolysis, and antioxidant systems, were identified. This has the potential to jeopardize the production of ATP (Adenosine Triphosphate) and the ability of erythrocytes to maintain their structures. In addition, they emphasize that oxidative stress is elevated as a result of the accumulation of reactive oxygen species, causing damage to the lipids and proteins of the cell membrane. This damage caused erythrocytes to be more fragile and to produce hemolytic processes; likewise, they propose that these metabolic transformations could change the deformability of erythrocytes, since it is a fundamental element for their transit through capillaries. In other words, they suggest that the spatial environment not only decreases the number of erythrocytes but also affects their quality and functioning [17].

Therefore, evidence suggests that erythrocyte fragility is closely related to physiological stress conditions. Thus, a study carried out by the authors showed that even mild systemic inflammation can significantly increase the osmotic fragility of erythrocytes. The authors reported that inflammatory processes alter the structural integrity of the erythrocyte membrane. Likewise, the authors reported that inflammatory processes alter the structural integrity of the erythrocyte membrane, making red blood cells more susceptible to hemolysis, this finding is very pertinent, because it reveals how barely perceptible changes in the internal environment can condition the stability of erythrocytes. As a consequence, the results obtained also suggest that external factors such as microgravity can cause similar alterations in the membrane and increase the sensitivity of erythrocytes in response to osmotic stress.

B. Theoretical Basis

1) Erythrocytes: Erythrocytes are specialized blood cells whose main function is the transport of oxygen from the lungs to the tissues and the return of carbon dioxide. Therefore, their biconvex structure allows them a high capacity for formation, facilitating their passage through narrow capillaries; therefore, the integrity of their cell membrane is essential to maintain their functionality and survival in the bloodstream [18].

2) Erythrocyte membrane: The erythrocyte membrane is composed of a lipid bilayer and structural proteins that ensure its flexibility, stability, and mechanical adaptability. Alterations in their composition, either due to oxidative stress or external factors, can affect cell deformability and increase the probability of hemolysis [19].

3) Osmotic fragility of erythrocytes: Osmotic fragility is a parameter that assesses the resistance of erythrocytes to changes in the osmotic concentration of the medium. Therefore, it is measured by observing the degree of hemolysis when cells are exposed to hypotonic solutions, i.e., an increase in osmotic fragility indicates a lower stability of the erythrocyte membrane, which implies greater susceptibility to cell rupture [20].

4) Microgravity: Microgravity is a physical condition characterized by a significant reduction in gravitational force; in this sense, it occurs in space or in simulated laboratory models. This condition generates physiological changes in the human body, including fluid redistribution, metabolic alterations, and modifications in cellular function, especially in the hematological system [21].

5) Oxidative stress: Oxidative stress occurs when there is an imbalance between the generation of reactive oxygen species and the antioxidant capacity of the body. In erythrocytes, this phenomenon can cause damage to membrane lipids and proteins, compromising their stability and increasing their fragility [22].

6) Hemolysis: Hemolysis is the process of destruction of erythrocytes, which can occur due to mechanical, chemical, or physiological factors. In conditions such as microgravity, an increase in hemolysis has been observed, which contributes to the decrease in erythrocyte mass and the development of anemia [23].

7) Erythrocyte deformability: Erythrocyte deformability is the ability of red blood cells to change shape without breaking through capillaries. This property depends on the integrity of the membrane and the cell cytoskeleton. Likewise, its alteration

can affect blood circulation and promote hemolytic processes [24].

III. Materials and Methods

In order to analyze in a structured way the scientific evidence regarding the effects of simulated microgravity on the osmotic fragility of erythrocytes, the present work was carried out under a systematic review of the literature. This methodology was chosen because of the way in which it evaluates, organizes and synthesizes the scientific information it has, in a rigorous and reproducible way and in which the minimum bias in the selection of studies intervenes. It should be noted that the prism methodology was used as a set of guidelines to be followed in the preparation and presentations of meta-analysis and systematic review, with the main purpose of this method being to ensure that the search and selection treatment of studies is done in a transparent and orderly manner. Likewise, the methodology allows to describe in detail how the search for the information has been done, as well as the criteria that were used to include/exclude the studies that were retrieved and which studies were finally analyzed, since it includes a flow chart that allows you to see the entire process of selecting the articles. In this way, the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology increases the reliability and quality of systematic reviews and meta-analyses, since other researchers can understand and replicate the procedure followed [25].

The bibliometric analysis was carried out in order to identify trends, research patterns and scientific collaboration networks in the study of spatial anemia and erythrocyte alterations under microgravity conditions. To this end, the VOSviewer software was used, which allowed the generation of graphical representations of keyword co-occurrence, co-authorship between researchers and collaboration between countries, using bibliographic data obtained from Scopus, IEEE Xplore, PubMed and Web of Science (WoS). This analysis facilitated the identification of the main lines of research and the structure of scientific production in the area [26].

A. Study Design

The research was based on a process of research and analysis of scientific research related to the effect of microgravity on hematological parameters, especially in terms of the stability and osmotic fragility of erythrocytes. Decision and experimental studies that focused on microgravity models, such as the clinostat or other simulation devices, were also noted.

B. Data Sources

In order to identify the relevant articles, an exhaustive bibliographic search was carried out in scientific databases of great international recognition, such as ScienceDirect, Scopus, PubMed, IEEE Xplore, which were selected for the bibliographic search due to the high level of coverage in the biomedical field for the indexing of scientific journals also of high impact.

C. Search Strategy

To search for the relevant studies, a combination of keywords such as microgravity, erythrocytes and osmotic fragility was carried out, in order to have a better coverage, synonyms and alternatives of the most frequent terms in the scientific literature were used. To this end, the four databases most used in the various scientific studies were used, such as WoS, Scopus, PubMed and IEEE. In addition, studies were collected during the period from 2021 to 2025, to obtain the most recent information from the literature and have a better outlook. For this purpose, Boolean operators such as (AND, OR) were used in the aforementioned databases. The strategies used are detailed below:

1) Web of Science: TS=((microgravity OR weightlessness OR "spaceflight" OR "simulated microgravity") AND (erythrocyte* OR "red blood cell*" OR RBC) AND (anemia OR hemolysis OR "red cell turnover" OR "osmotic fragility" OR "membrane stability" OR deformability))

2) Scopus: TITLE-ABS-KEY((microgravity OR weightlessness OR "spaceflight" OR "simulated microgravity") AND (erythrocyte* OR "red blood cell*" OR RBC) AND (anemia OR hemolysis OR "red cell turnover" OR "osmotic fragility" OR "membrane stability" OR deformability))

3) PubMed: ((microgravity[Title/Abstract] OR "simulated microgravity"[Title/Abstract] OR spaceflight[Title/Abstract] weightlessness[Title/Abstract]) AND (erythrocyte*[Title/Abstract] OR "red blood cell*"[Title/Abstract] OR RBC[Title/Abstract]) AND (anemia[Title/Abstract] OR hemolysis[Title/Abstract] OR "red cell turnover"[Title/Abstract] OR "osmotic fragility"[Title/Abstract] OR "membrane stability"[Title/Abstract] OR deformability[Title/Abstract] OR "space anemia"[Title/Abstract])) AND (ffrft[Filter]) AND ("2021/01/01"[Date -Publication] : "2025/12/31"[Date - Publication])

4) IEEE Xplore: (("Publication Title": microgravity OR "Publication Title": "simulated microgravity" OR "Publication Title": spaceflight OR "Publication Title": weightlessness) AND ("Abstract": erythrocyte OR "Abstract": "red blood cells" OR "Abstract": RBC) AND ("Abstract": anemia OR "Abstract": hemolysis OR "Abstract": "red cell turnover" OR "Abstract": "osmotic fragility" OR "Abstract": "membrane stability" OR "Abstract":d eformability))

D. Inclusion and Exclusion Criteria

In order to ensure the validity, relevance, and updating of the works that are recognized, specific inclusion and exclusion criteria have been determined.

- Inclusion criteria:
  - Articles published in indexed journals.
  - Scientific articles published in the last 5 years.
  - Publication in English.
  - Investigations that include analysis of osmotic fragility, erythrocyte membrane stability or hemolysis.
  - Investigations that evaluate erythrocytes, spatial anemia or hematological parameters.

- Investigations related to simulated or real microgravity.
- Studies that use experimental models of microgravity, such as the clinostat.

- Exclusion criteria:
  - Studies with insufficient information or unclear methodologies.
  - Studies without access to full text.
  - Duplicate Articles Across Databases.
  - Publications that do not address the specified topics.

E. PRISMA Methodology

The strategy used for the selection of the review articles presented below was the one found in the protocol, which made possible the orderly and reproducible structures of each of the review phases. Therefore, in the identification phase, the search for these articles from the databases and using the Boolean operators mentioned above was carried out. As can be seen in Fig. 1, in the first instance in the Scopus database, a number of 41 articles were obtained, in IEEE Xplore 178, in PubMed 200 and in Web of Science 228, obtaining a total of 745 articles to be later analyzed. Then, in the filtering phase, duplicate records and those that did not meet the inclusion and exclusion criteria were eliminated, such as those that were not related to the topic of the study, year of publication, as well as those that were not available in full text, leaving a total of 75 articles.

On the other hand, in the screening phase, another evaluation was carried out, but this time the titles and abstracts of the articles were verified, according to the direct relationship, taking into consideration the subject of study in this specific case, in order to determine the effect of microgravity on the behavior of erythrocytes, leaving a total of 42 articles. Therefore, in the inclusion stage, a conscientious reading of the differently selected articles was carried out, where duplication was found in two articles; therefore, a total of 40 articles were finally left. In this phase, its scientific relevance and methodological relevance were critically assessed, also adding that the reading of the articles from which the set of studies carried out subsequently for qualitative research was ultimately elaborated, as well as in short, this procedure allowed to gradually reduce the number of studies initially selected and also to guarantee the selection of appropriate and updated research related to qualitative research line of the objectives of this study [27].

F. Analysis of Information

In this section, the research selection process that underpins the systematic review that was carried out according to the PRISMA diagram model is detailed, in which the phases of identification, filtering, screening, and inclusion of the studies formed in a logical sequence are organized. In the first stage of identification, a total of 745 of the records were obtained from the 4 selected databases and were collected: Scopus (41), IEE Xplore (178), PubMed (200), Web of Science (226). In this sense, this phase aimed to obtain a high number of publications that could be relevant in terms of microgravity and its effects on erythrocytes in order to configure a broad initial base for the analysis, Thus, in the second phase of filtering, the inclusion and exclusion criteria were established in order to purify the records obtained, Likewise, in this phase, criteria were established such as the type of document taken as a basis, with preferable those investigations that are aligned with the original article, as well as the direct relationship with the topic of study pointed out in terms of the conditions of microgravity or spaceflight, anemia, impact on erythrocytes and their implications.

However, specific variables such as hemolysis, anemia, osmotic fragility, membrane stability, and deformability were evaluated. It should be noted that only studies published in the period of 2021-2025 were included, corresponding to articles available in English and, likewise, with access to the full text. As a result of this process, around 75 articles were obtained in the eligibility phase: Scopus (28); IEEE Xplore (24); PubMed (4); Web of Science (19). In the screening phase, the titles and abstracts of the articles obtained were meticulously analyzed. In this phase, 42 investigations were obtained that were taken into consideration for subsequent analysis. In this phase, 33 studies were excluded since they did not meet the methodological or thematic criteria that we had established. Therefore, in the inclusion phase, only 2 duplicate articles were detected; therefore, they were eliminated to avoid content redundancies. Therefore, a set of 40 articles was obtained that make up the database of this systematic review, which were organized systematically in order to facilitate the analysis.

For each collection study, information was collected based on relevant bibliographic data, such as the year of publication, authors, source of origin, which allowed it to be structured in an orderly manner, facilitating the analysis of the different approaches used in research on erythrocyte alterations in conditions of microgravity, spatial anemia, as well as the possibility of contrasting the methods in the different articles. However, for the organization and management of the articles, specialized tools such as bibliographic managers were used, in which it was very useful, since they allowed storing the records, identifying duplicate articles and thus optimizing the review process, likewise, the VOSviewer software was used for the elaboration of bibliometric maps and collaboration networks between authors and countries, allowing a visual analysis of scientific production to be carried out.

Together, the selection method applied to the literature made it possible to reduce a large amount of scientific literature until a final set of relevant studies was obtained. It should be noted that the rigorous application of the inclusion and exclusion criteria in each phase made it possible to guarantee the coherence of the selected articles, building a solid database for the analysis of erythrocyte alterations under microgravity conditions and the analysis of spatial anemia. In this way, the included studies provide an up-to-date view of the state of research in this field, as well as a starting point for future scientific research (see Fig. 1).

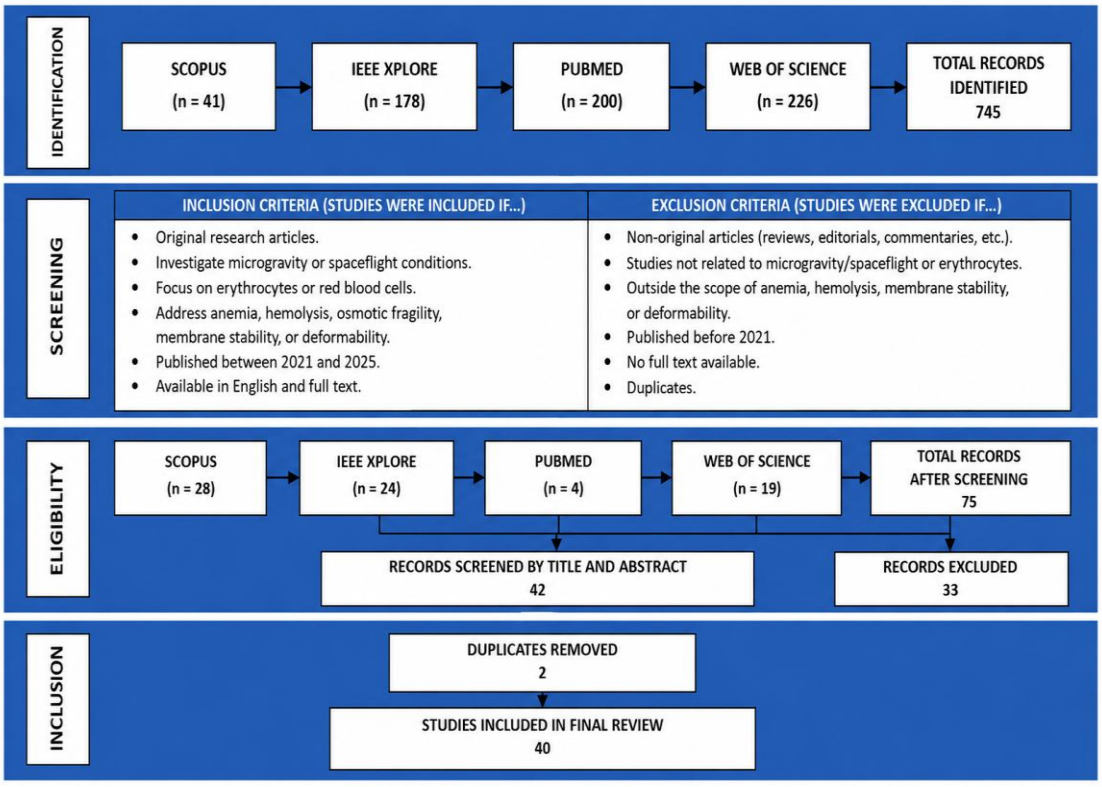


Fig. 1. PRISMA methodology

IV. Results

In this section are the results of the systematic review. There was an increase in scientific interest in researching the effects of microgravity on erythrocytes, with a particular focus on problems such as hemolysis, spatial anemia and alterations in the mechanical characteristics of red blood cells. In this sense, the information based on the analyzed literature shows how this field has been consolidating itself as a relevant area within aerospace biomedical research, due to its impact on the health of astronauts during their prolonged missions.

In general terms, the studies reviewed show an inclination towards the use of experimental models, both in real and simulated microgravity conditions, as well as including places such as the International Space Station and devices designed for simulation on Earth. These approaches allow the analysis of physiological changes in erythrocytes such as osmotic fragility, membrane stability, deformability and in general the fundamental aspects in order to understand the appearance of the so-called spatial anemia, in this way, the results allowed to identify variations in the methods used, as well as the type of samples analyzed and the level of evidence achieved by the studies. An unequal geographical distribution of scientific production is also observed, these findings offer a current view of the state of research and serve as the basis for the bibliometric analysis developed below.

A. Keyword Co-occurrence Map

In this section, an analysis of keyword co-occurrence was carried out with the 40 selected articles, using the VOSviewer tool, with the aim of determining the most important thematic lines in research on microgravity and erythrocytes. Therefore, in Fig. 2, you can see the map produced where he identified an extensive network of interrelated terms, organized in multiple clusters. In this sense, in this representation, the size of nodes indicates the frequency of occurrence of each keyword, however, the thickness of the links reflects the intensity of the relationship between the terms and the colors, since they facilitate the distinction of different thematic groups [28].

Likewise, several central nuclei were identified that are clearly delimited. Therefore, the concepts "spaceflight" "microgravity", "erythrocyte", and "hemolysis" are the ones that stand out the most, which constitute the main axis of the research, and demonstrate that most of the studies are carried out in space conditions or simulated microgravity. Likewise, terms such as "ISS" or "international space station" emphasize the importance of the experimental spaces used in this field. In addition, a group focused on "RBC", "erythrocyte", and "erythrocytes" is seen. Therefore, this confirms that erythrocytes are the most important biological center of the literature analyzed. This set is strongly linked to physiological variables such as hemolysis, which is presented as one of the most essential terms and thus shows interest in the process of damage or destruction of red blood cells in microgravity conditions.

In this sense, terms related to biochemical and cellular processes are identified, such as "function", "oxidative stress", "membrane", "lipidomic", which indicate that a significant part of the research is focused on understanding the underlying mechanisms that explain the functional and structural alterations of erythrocytes [29]. Therefore, the presence of terms related to

methodological and technological aspects was identified, such as "flow", "ultrasound", "impedance", "channel", which indicated the use of various experimental techniques for the analysis of erythrocyte properties [30].

Overall, the co-occurrence map shows a strong relationship between microgravity conditions and the changes they cause in erythrocytes, particularly hemolysis and oxidative stress, which are essential processes for understanding spatial anemia [31]. It also demonstrates the combination of experimental and technological methods in investigating this phenomenon, which highlights the multidimensionality and complexity of research in this area (see Fig. 2).

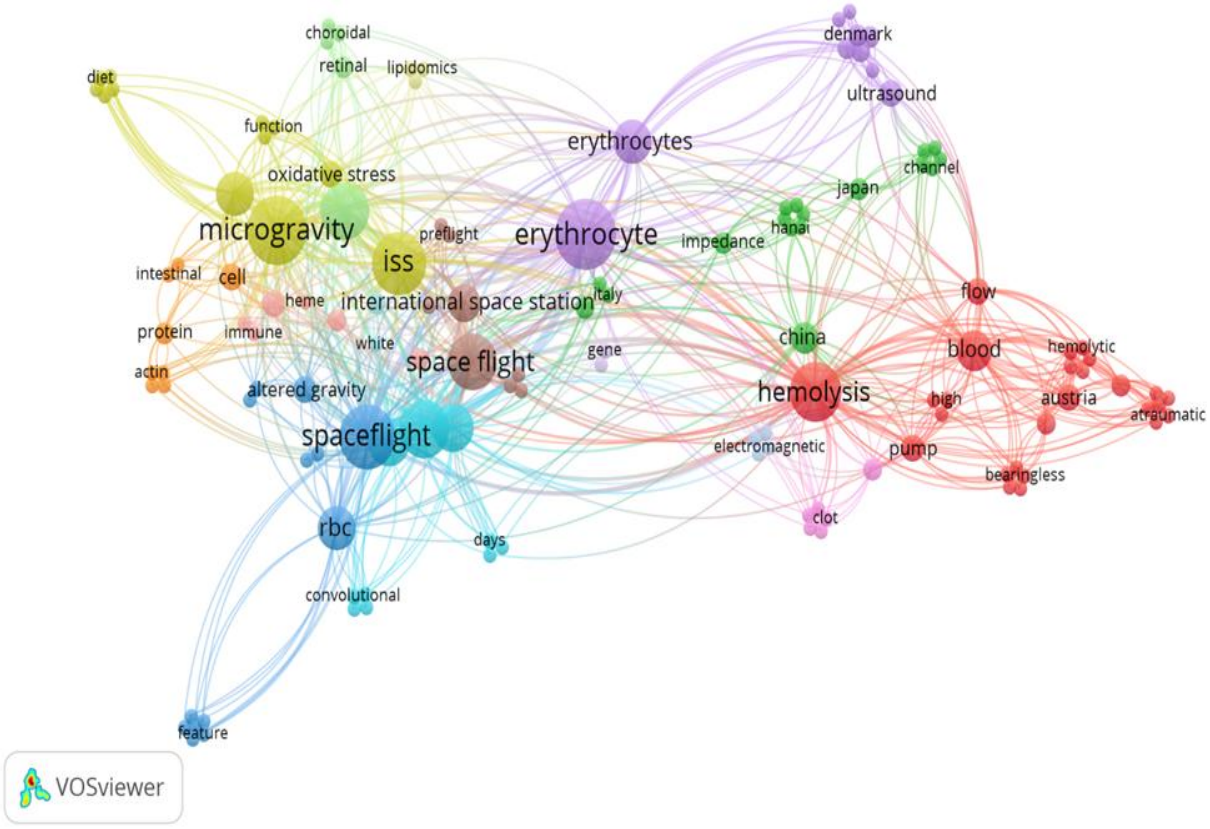


Fig. 2. Keyword co-occurrence map

B. Co-authorship Map

A co-authorship analysis was carried out with the aim of determining the cooperative links between scientists who have contributed to the study of microgravity and its repercussions on erythrocytes. In this way, the VOSviewer software was used, which made it possible to graphically represent the relationships between authors from the 40 articles chosen, as can be seen in Fig. 3; therefore, the map obtained reveals a network organized into two main groups of collaboration. The first largest cluster brings together authors such as Ryan, K. A., Houerbi, N., Kim, J., Mason, C. E., Mateus, J., Najjar, D., Afshin, E. E., and Damle, N., all of whom have numerous connections between them.

Therefore, the high concentration of links shows a close and recurrent collaboration within this group, which indicates the presence of a consolidated research team working in this thematic line [32]. On the other hand, a second cluster formed by Overbey, E. G., and Pecaut, M. J., characterized by a smaller structure and with fewer connections, is identified. However, the role of Overbey, E. G., as a link node between both groups stands out, since it maintains connections both with the main cluster and with other authors, thus facilitating interaction between different lines of research.

In this way, the presence of these two groups reflects a partially fragmented network of collaboration, in which highly interconnected research nuclei coexist with other, more independent ones. Likewise, the existence of authors who act as bridges between clusters highlights the importance of certain researchers in the articulation of knowledge within the field. Overall, the co-authorship analysis suggests that, although there are consolidated research groups in the study of microgravity and erythrocytes, global collaboration is still at a moderate level, which opens the possibility of strengthening broader scientific networks in future research (see Fig. 3).

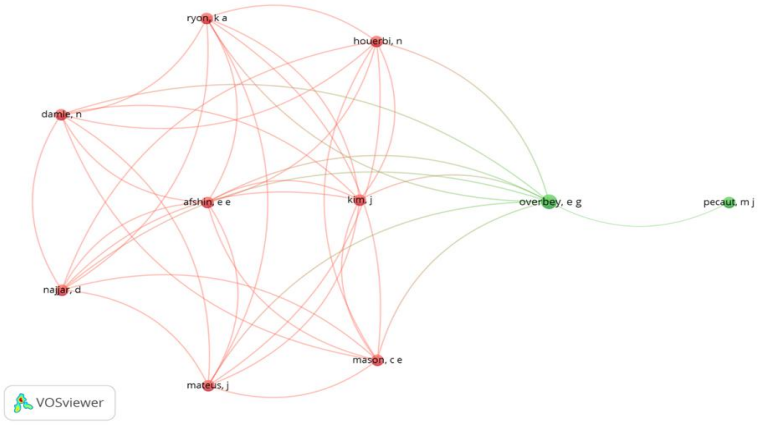


Fig. 3. Co-authorship map

C. Map of Co-authorship by Country

In this section, the graphical representation facilitates the analysis of international collaboration between countries within a set of selected studies, as can be seen in Fig. 4. Two main groups were identified that were distributed in five countries and are linked by several co-authorship links. Thus, in this type of map, each node represents a country in the same way the lines indicate the existence of publications made by a group of authors from different nations [33].

In this sense, the network shows that international collaboration is limited, due to the fact that there are few countries that have co-authorship links, which indicates that most of the research is carried out in small teams and is also developed in national contexts. However, Austria is a point of connection between the European group (Austria, Germany, Switzerland) and the Asian group (China and Japan), allowing for intercontinental interaction [34]. On the other hand, closer ties are observed in each of the groups, especially in the European bloc, indicating more consolidated regional cooperation. In Asia, cooperation and collaboration are mainly concentrated in Japan and China, the latter being an important node because it maintains relations not only within its geographical area but also with Europe; therefore, the analysis of co-authorship reveals a sparse network with concrete links, which evidences a specialized field of research. This indicates that, despite international cooperation efforts, there is still room to strengthen cooperation between countries and to promote joint research that broadens the scope and also the robustness of research in this area (see Fig. 4).

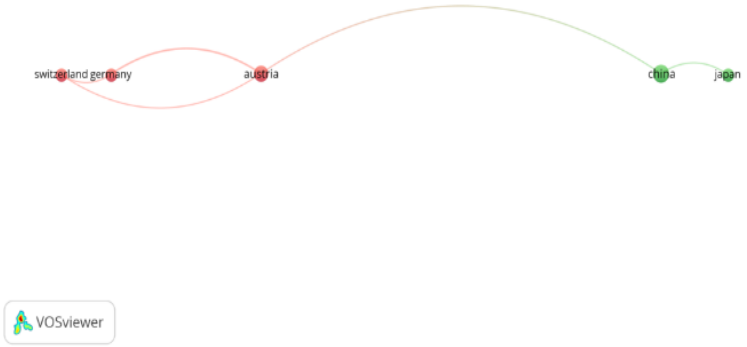


Fig. 4. Co-authorship map by country.

D. Annual Evaluation of Scientific Production

Fig. 5 shows how the annual scientific production linked to research into microgravity and its impacts on erythrocytes has changed. The review was carried out based on the 40 selected articles. Thus, the vertical axis of the graph shows how many articles were published in each period, while the horizontal axis shows the representation of the years of publication, in general terms, between 2021 and 2025. Scientific production represents a variable behavior. Consequently, 9 publications were registered in 2021, but the production dropped to 7 articles in 2022. Subsequently, in 2023, the lowest value of the entire period (3 publications) is appreciated. This indicates a temporary decrease in the number of studies reported on this topic.

Since 2024, a significant increase in scientific production has been observed, reaching its peak in the period with 12 investigations. This increase indicates a more recent interest in investigating alterations in erythrocytes that occur in microgravity, such as spatial anemia, hemolysis, and variations in red blood cell function and stability. In 2025, production remains at a high level with 9 articles, which supports the permanence of this line of research. In general, the trend observed despite the fact that production was not constant during all years, in the most recent periods there was a significant increase. This allows us to infer that research on microgravity and its impacts on erythrocytes has established itself as an emerging area of interest, with a projection to continue developing in future research [35].

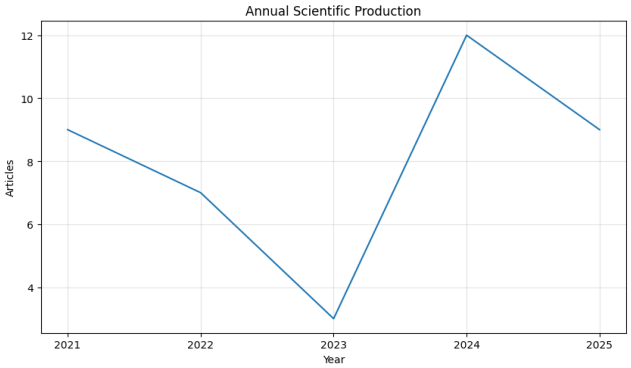


Fig. 5. Annual scientific production

E. Distribution of Publications by Journal

This section refers to the distribution of publications by scientific journals, as can be seen in Fig. 6. It allowed us to identify the most relevant sources within the field of study related to microgravity and erythrocytes. Therefore, with the results obtained from 40 articles, it was identified that the journal Studies in Health Technology and Informatics concentrates the largest number of publications, with a total of 5 articles, positioning itself as the main source of dissemination in this field. Therefore, this result indicates that it is of utmost importance in research that integrates technology and health, especially in innovative contexts such as microgravity.

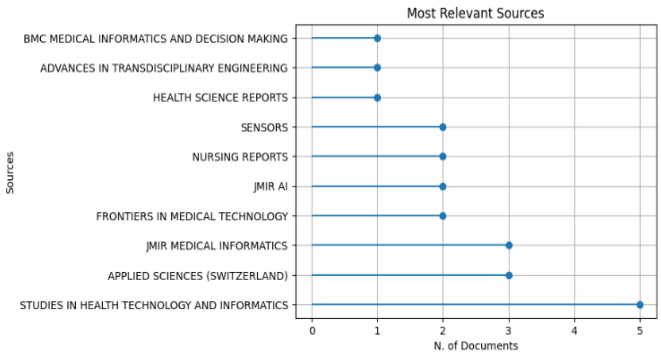


Fig. 6. Most relevant sources

Consequently, the journals Applied Sciences (Switzerland) and JMIR Medical Informatics stand out, with 3 publications each, which shows their significant contribution to the dissemination of studies related to computational and technological analysis applied to biological phenomena. Likewise, several journals with an intermediate production are identified, such as Frontiers in Medical Technology, JMIR AI, Nursing Reports and Sensors, each with 2 published articles. These sources reflect the thematic diversity of the field, spanning approaches from biomedical engineering to clinical applications.

On the other hand, journals such as Health Science Reports, Advances in Transdisciplinary Engineering and BMC Medical Informatics and Decision Making have a lower participation, with 1 article each, which indicates a more specific contribution within the analyzed set. Overall, the observed distribution shows that scientific production is not concentrated in a single journal, but is dispersed in various sources, which confirms the interdisciplinary nature of the research topic. In addition, this diversity favors the dissemination of knowledge and the development of new lines of study in the field of microgravity and its effects on erythrocytes [36].

F. Content Analysis

With respect to content analysis, the selected articles allowed the identification of the main methodological, thematic and conceptual approaches present in the literature in relation to microgravity and its effects on erythrocytes. Therefore, this analysis was carried out based on the review of the 40 studies, which made it possible to organize the information into important categories. It should be noted that a significant part of the research was identified focused on the study of the physiological alterations of erythrocytes under the conditions of both real and simulated microgravity. In this sense, the scientific articles analyzed highlight changes in the stability, structure and functionality of red blood cells, evidencing their impact on processes such as spatial anemia and hemolysis. Therefore, with respect to the analysis of the osmotic fragility of erythrocytes, it is considered as a key indicator of the integrity of the cell membrane. Therefore, several studies agree that exposure to microgravity can increase the susceptibility of erythrocytes to rupture, which causes an alteration in their mechanical and biochemical properties [37].

On the other hand, research aimed at the study of molecular and cellular mechanisms, including oxidative stress, is identified; therefore, the factors are frequently related to the decrease in erythrocyte stability and the increase in hemolytic processes. As for the methodological approaches, the use of experimental models of simulated microgravity, such as the clinostat, as well as studies carried out in real environments such as the International Space Station, is evidenced. Therefore, these methods make it possible to analyze the effects of the space environment on blood cells in a controlled manner. Likewise, the content analysis identified that the studies present a diversity of interdisciplinary approaches, thus integrating areas such as biomedicine, biochemistry, physiology, and engineering. In this way, this reflects the difficulty of the phenomenon studied and also the need to address it from different perspectives. Thus, the results regarding the content analysis showed that research in this field of development has made important advances in the understanding of the effects on microgravity on erythrocytes. However, gaps in the literature are also identified, especially in relation to long-term clinical studies and the standardization of methodologies, which suggests the need to continue deepening this line of research [38].

G. Distribution of Journals on Microgravity and Erythrocyte Studies

Fig. 7 shows the distribution of the scientific articles analyzed according to the journals in which they were published, which allows us to identify the main sources of dissemination when investigating microgravity and its impacts on erythrocytes. Therefore, from the 40 articles selected in this systematic review, it is observed that the journal Studies in health Technology and Technology and Informatics concentrates the largest number of publications with a total of 5 articles, positioning itself as the most relevant source in the field. Thus, this result shows the focus on research that integrates

health and technology, especially in innovative contexts such as microgravity. It should also be noted that the journals (Switzerland), JMIR Medical Informatics and Applied Sciences with 3 publications each, reflect their significant contribution to the dissemination of studies related to the analysis of biological phenomena through computational and technological approaches.

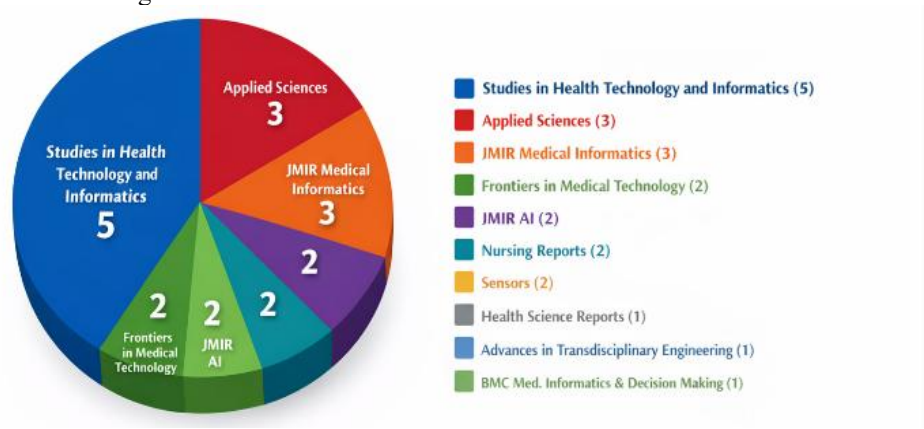


Fig. 7. Content analysis

Likewise, several journals with an intermediate production are identified, including Frontiers in Medical Technology, JMIR AI, Nursing Reports and Sensors, each with 2 published articles. These journals reflect the diversity of approaches present in the literature, ranging from biomedical engineering to clinical applications related to the analysis of erythrocytes in microgravity conditions. On the other hand, journals such as Health Science Reports, Advances in Transdisciplinary Engineering and BMC Medical Informatics and Decision Making have a lower participation, with 1 article each, which indicates a more specific contribution within the analyzed set.

Thus, the observed distribution shows that the scientific production is dispersed in various journals, which confirms the interdisciplinary nature of the study of erythrocytes and microgravity. Therefore, this diversity favors the dissemination of knowledge and also contributes to the development of new lines of research in the field of biochemistry and aerospace medicine [39].

H. Comparison of Studies on the Effects of Microgravity on Erythrocytes and Their Relationship with Osmotic Fragility

As can be seen in Table I, of the 40 articles that were selected, only 8 studies were taken into account for a comparative analysis, because only this small group of articles provided quantitative or qualitative information directly associated with erythrocyte osmotic fragility, hemolysis, membrane stability, or equivalent indicators associated with the adaptation of erythrocytes to microgravity conditions. Likewise, these studies had a more complete description of the experimental protocols used and reported results that facilitated comparisons between the different methodological approaches. It should be mentioned that the remaining studies were considered during the general synthesis of the systematic review evidence, but did not provide sufficient information about the specific variables analyzed in this table. To this end, they were distributed by authorship, year, title, experiment-protocol (brief), metrics (main findings) and limitations, likewise, it is evident that most of the studies analyzed coincide in microgravity conditions, both simulated and real, which can generate important alterations in erythrocytes. Therefore, these modifications are not always directly evaluated by classic tests such as the OFT (Osmotic Fragility Test) or MCF/H50 (Mean Corpuscular Fragility-H50) tests, but they do manifest themselves through functional, biochemical and structural changes that affect cell stability. Therefore, firstly, some studies, such as that of Agramont et al. (2023), directly report an increase in osmotic fragility and also in susceptibility to hemolysis, suggesting a lower resistance of erythrocytes to changes in the osmotic environment. This type of evidence is especially relevant, as it allows a more direct relationship to be established between erythrocyte damage and microgravity.

Therefore, a large part of the studies provides indirect evidence, since research such as that of Manis (2022, 2023) indicates an increase in oxidative stress, alterations in the composition of the membrane and metabolic changes, factors that are closely related to the loss of cell stability. Therefore, Murali and Sarkar (2025) analyze changes in the deformability and behavior of erythrocytes, which may affect their ability to resist adverse situations and increase their susceptibility to hemolysis. In studies carried out in real microgravity, such as Trudel et al. (2022), there are evidence of a sustained increase in hemolysis during prolonged space missions, which supports the notion that the destruction of red blood cells is an active process in this context, which favors the emergence of the so-called spatial anemia. In this sense, according to Hinge et al. (2024) they show morphological and functional alterations in erythrocytes both under simulated conditions and in reality, which indicates that experimental models on the ground can partially replicate the effects seen in space. Finally, certain studies considered as additional support according to Stuart et al. (2024) and Kashirina et al. (2022), offer evidence from a physiological and molecular perspective. These studies indicate that elements such as inflammation or molecular stress may increase the fragility of erythrocytes, thus contributing to a better understanding of the underlying mechanisms, even if they are not directly related to microgravity.

TABLE I. PROPOSED MODEL STUDIES

No.	Author	Year	Title	Experiment-Protocol (brief)	Metrics - Main findings	Limitations
[9]	Agramont N.; Carrasco M.; Flores B.; Quenta D.; Chávez G.	2023	Osmotic Fragility and its Relationship with Space Anemia	A study focusing on the osmotic fragility of erythrocytes and its relationship to space anemia under conditions associated with simulated microgravity	Osmotic fragility: increased. Hemolysis: increased susceptibility. MCF/H50: not reported. p-value: not reported	The abstract does not provide standardized quantitative values; there is limited numerical evidence in the available excerpt.
[10]	Manis C.; Manca A.; Murgia A.; Uras G.; Caboni P.; Congiu T.; Faa G.; Pantaleo A.; Cao G.	2022	Understanding the Behavior of Human Cell Types under Simulated Microgravity Conditions: The Case of Erythrocytes	Human red blood cells under simulated microgravity; analysis of structural and biochemical changes in the cell membrane	Oxidative stress: increased. Antioxidant capacity: decreased. Erythrocyte membrane: altered. Osmotic fragility: indirect evidence of an increase. MCF/H50: not reported. p-value: not reported	It does not directly assess OFT or MCF50 in the available content; the relationship to osmotic fragility is indirect
[12]	Manis C.	2023	Study of metabolic changes in human erythrocytes and plasma under simulated microgravity conditions	Metabolic study of erythrocytes and plasma under simulated microgravity	Erythrocyte metabolic changes: present. Cell stability: compromised. Osmotic fragility: indirect evidence of increased levels. MCF/H50: not reported. p-value: not reported.	The abstract does not include complete numerical results or a full description of the methodology; verification against the full text is required.
[11]	Murali A.; Sarkar R. R.	2025	Impact of gravitational forces on Red Blood Cell dynamics in biofluid suspension	Evaluation of erythrocytes under different gravitational conditions; analysis of shape, deformability, and cell dynamics	Erythrocyte deformability: altered. Cell dynamics: altered. Susceptibility to hemolysis: indirectly increased. Osmotic fragility: indirect evidence. MCF/H50: not reported. p-value: not reported	It does not report direct measurements of osmotic fragility in the study; the evidence is primarily mechanical and functional.
[13]	Trudel G.; Shahin N.; Ramsay T.; Laneuville O.; Louati H.	2022	Hemolysis contributes to anemia during long-duration space flight	Study of actual microgravity during long-duration spaceflight; analysis of hemolysis and space anemia.	Hemolysis: consistently elevated. Space-occupying anemia: associated with persistent red blood cell destruction. Osmotic fragility: not directly assessed. MCF/H50: not reported. p-value: not reported in the abstract.	The primary outcome is hemolysis, not direct OFT; in a real-world setting, there may be additional confounding factors related to spaceflight.
[5]	Hinge S.; Dixit J.; Vidyasagar P.; Kulkarni G.	2024	Comparative analysis of red blood cell under real and simulated microgravity conditions	Comparison of erythrocytes in real and simulated microgravity; structural and functional analysis	Red blood cell morphology: abnormal. Functional changes: present. Osmotic fragility: not directly reported in the abstract. MCF/H50: not reported. p-value: not reported	The exact quantitative results are not detailed in the abstract; a full review of the article is required for a detailed analysis.
[2]	Kashirina D. N. et al.	2022	Proteomic characterization of dry blood spots of healthy women during simulation the microgravity effects using dry immersion	Simulated microgravity model using dry immersion; proteomic characterization of blood samples	Hematological proteomic changes: present. Osmotic fragility: indirect evidence; not measured directly. Hemolysis: not reported directly. MCF/H50: not reported. p-value: not reported in the abstract.	It provides additional molecular evidence, but does not focus directly on OFT or quantified erythrocyte hemolysis.
[17]	Stuart C. M.; Jacob C.; Varatharaj A.; Howard S.; Chouhan J. K.; Teeling J. L.; Galea I.	2024	Mild Systemic Inflammation Increases Erythrocyte Fragility	Clinical study of erythrocyte fragility in the context of mild systemic inflammation; included as a complementary mechanistic basis	Erythrocyte fragility: increased. Microgravity: not directly assessed. MCF/H50: not reported in the abstract. p-value: not reported in the abstract	This does not correspond to direct microgravity; it should be interpreted as supplementary evidence rather than as a key spatial finding.

Thus, the findings suggest that although there is an evident tendency to increase the fragility and hemolysis of red blood cells in situations of microgravity, the information found is diverse and in many cases is indirect, likewise, there is a lack of standardization of the methodologies used, since few studies present specific quantitative parameters such as MCF/H50 (Mean Corpuscular Fragility-H50) or osmotic fragility curves, which complicates the direct comparison between the investigations [40].

V. DISCUSSION

The results of this systematic review show that there is a general consensus that microgravity affects erythrocyte stability, however, the mechanisms responsible for these alterations differ among the studies analyzed. While some studies attribute these changes mainly to membrane damage and oxidative stress, others suggest that metabolic alterations and reduced cell deformability play a relevant role. These differences may be related to the variability of the experimental models used and the duration of exposure to microgravity. It should be noted that, in

both simulated and real situations, they have a significant impact on the functional and structural stability of erythrocytes. Therefore, the research examined indicates that this environment favors alterations associated with hemolysis, membrane damage, metabolic variations and decreased cell deformability, thus all the processes related to the emergence of special anemia and a possible increase in the osmotic fragility of red blood cells. Although not all studies directly analyzed classical parameters such as OFT (Osmotic Fragility Test) curves or MCF/H50 (Mean Corpuscular Fragility-H50) values, the evidence collected suggests a clear trend towards deterioration of erythrocyte integrity when exposed to microgravity [41].

Overall, the results obtained support the notion that microgravity affects the red blood cell membrane and thus alters its ability to withstand variations in the osmotic medium. This is most clearly evidenced in the research of Agramont et al. [10], in which osmotic fragility is linked with spatial anemia and a greater predisposition to hemolysis. This discovery is particularly significant as it establishes a more direct relationship between microgravity conditions and reduced cell stability. Consequently, from a pathophysiological approach, this indicates that exposure to extreme situations can weaken the mechanical and osmotic resistance of the erythrocyte, in this sense, favoring its Destruction and altering the interpretation of hematological analyses in aerospace environments. However, this finding coincides with the results reported by Trudel et al., who also identified an increase in erythrocyte destruction during prolonged exposures to the space environment; however, Agramont et al. establish a more direct relationship related to osmotic fragility, while Trudel et al. highlight hemolysis as the main clinical manifestation. This difference reflects the diversity of approaches used to study spatial anemia.

However, several studies incorporated in this review presented indirect but consistent evidence of red blood cell deterioration. Therefore, studies by Manis et al. [13] showed that an increase in oxidative stress, modifications in the lipid composition of the membrane, reduction of antioxidant capacity, and alterations that directly impact the stability of red blood cells and homeostasis. Similarly, Murali and Sarkar [12] analyzed changes in the shape, deformability, and dynamics of erythrocytes under various gravitational conditions, emphasizing that gravitational field variation affects the physical properties crucial to erythrocyte survival. However, although these studies do not measure osmotic fragility uniformly through a standardized parameter, they reinforce the idea that microgravity causes changes that favor hemolysis and the loss of membrane integrity. Overall, these studies suggest that erythrocyte damage induced by microgravity is multifactorial. On the contrary, the absence of standardized methodologies makes it difficult to establish which mechanism has the greatest influence on osmotic fragility. In short, this limitation makes it impossible to determine whether oxidative stress is the primary cause of damage or a secondary consequence of other cellular alterations.

Therefore, research conducted in microgravity conditions provides an extra level of clinical significance. Trudel et al. [14] documented that hemolysis plays an ongoing role in anemia during prolonged spaceflight, indicating that erythrocyte destruction is not just a transient adaptive response, but a long-lasting active process. Next, the result challenges the traditional hypothesis that spatial anemia is only a transient adaptive response to the redistribution of body fluids. Meanwhile, the persistence of observed hemolysis suggests the existence of active biological mechanisms capable of compromising erythrocyte survival during prolonged missions. Therefore, this finding supports the hypothesis that the spatial environment influences erythrocyte longevity through various pathways, including oxidative stress, metabolic changes, and cellular membrane variations. In summary, spatial anemia should be considered a multi-factor phenomenon, in which the fragility of the erythrocyte plays a major role along with other system adaptation mechanisms.

It is worth highlighting that a crucial aspect that emerges from this review is the methodological variability of the studies analyzed, studies carried out in real microgravity, simulated models such as clinostat or dry immersion and molecular or biochemical approaches aimed at detailing alterations in membrane, proteomics in blood, or metabolism. This variety shows that the analysis of the erythrocyte in microgravity is a clearly interdisciplinary area, but at the same time highlights one of its main weaknesses: the lack of standardization, while certain studies examine the hemolysis, or others focus on deformability. Likewise, some prioritize molecular biomarkers and only a small fraction focus directly on osmotic fragility. This diversity complicates comparison between studies and thus restricts the ability to reach quantitative conclusions that are more important [42]. Importantly, not all studies reported the same magnitude of erythrocyte alteration. However, some studies documented significant changes in hemolysis, membrane stability, and deformability; others found more moderate effects. Consequently, these discrepancies can be attributed to differences in the experimental models used, as well as the exposure conditions, the assessment methods used and the exposure conditions, the duration. Therefore, direct comparison between studies should be made with caution.

From this perspective, one of the discoveries that is significant of this review is that most of the existing studies offer indirect evidence rather than direct measurements of osmotic fragility. This does not dismiss their contributions, but it does require interpreting the results with caution. By entering the modification of the erythrocyte membrane, the increase in oxidative stress, the existence of damaged lipids and also the alterations in energy metabolism are indications of greater cellular vulnerability, although it does not always correspond to an accurate diagnosis through standardized osmotic tests. Therefore, the findings of this review indicate that there is a strong biological basis linking microgravity to increased erythrocyte susceptibility, although research assessing this phenomenon in a direct and comparable way is still needed.

Similarly, the data analyzed indicate that microgravity influences not only the structure of the erythrocyte but also its physiological function. Therefore, decrease in the mass of red blood cells, biochemical alterations, and an increase in hemolysis. Findings in several studies support the notion that oxygen transport can be affected during prolonged exposures to the space environment. This has significant relevance for aerospace medicine, as it indicates that hematological

surveillance in astronauts must take into account not only the erythrocyte count, but also their functionality, stability, and vulnerability to destruction. That is, erythrocyte research in microgravity should not focus solely on quantitative analysis, but also include mechanical, biochemical and osmotic factors to understand more fully the phenomenon of anemia in space [43].

In addition, this review facilitates the identification of obvious gaps in the literature; there are few studies that present standardized quantitative values on osmotic fragility MCF-H50 (Mean Corpuscular Fragility-H50), OFT (Osmotic Fragility Test) curves, or percentages of hemolysis under similar protocols. Therefore, numerous studies show small samples, as well as varied experimental designs or methods focused on indirect variables. However, research remains fragmented among several disciplined models, which complicates the creation of a coherent framework of interpretation. These restrictions indicate that, although studies have increased considerably in recent years, still there is not enough methodological uniformity to make direct comparisons solid between investigations.

Thus, the trend presented in this review is Coherent: Microgravity promotes changes in erythrocytes that increase cellular susceptibility and contribute to hemolytic processes linked to spatial anemia. However, the evidence you have found is stronger. Still, in the case of structural, functional and metabolic markers rather than direct osmotic evidence, the key contribution of this analysis focuses on bridging the scattered evidence and demonstrating that osmotic fragility can be seen as an important explanatory axis in erythrocyte pathophysiology in microgravity, although it is important to understand the importance of the Osmotic Pathophysiology. They need studies to be more concrete to validate it more accurately.

In other words, these findings suggest that the hematological monitoring of astronauts should incorporate additional functional indicators beyond the conventional erythrocyte count. That is, the evaluation of biomarkers related to hemolysis, membrane stability, and oxidative stress could contribute to an earlier detection of hematological alterations during prolonged space missions. In addition to this, the findings of this review indicate that future research should focus on creating experimental studies with standardized protocols, samples more extensive and also evaluations simultaneous of hemolysis, deformability, biomarkers of membrane damage, and osmotic fragility. A strategy of this type would make it easier not only to compare between studies, but also to enhance the understanding of the mechanisms that cause spatial anemia. Therefore, the osmotic fragility of erythrocytes is presented as an area of attractive interest for biomedical research and aerospace medicine, as it can be a marker of functional adaptation and cell damage under microgravity conditions.

VI. Conclusion

It is concluded, according to the analysis carried out, that microgravity in both simulated and real environments causes significant alterations in erythrocytes that impact functional and structural stability. Therefore, these variations are mainly expressed by an increase in hemolysis, as well as alterations in the cell membrane and changes in oxidative and metabolic processes that together favor the emergence of the so-called spatial anemia. Thus, the evidence analyzed in detail indicates that there is a connection between these alterations and a possible increase in the osmotic fragility of red blood cells. However, this connection has not always been examined directly, since most of the research reviewed is based on indirect indicators such as cellular elasticity, oxidative stress, or the lipid composition of the membrane.

In this sense, despite the constant phenomenon, there is still not enough research that applies standardized methods to evaluate osmotic fragility in a specific way. On the other hand, considerable variability was detected in the methods used, covering research in true microgravity, molecular analysis, and simulated models, which manifests the interdisciplinary nature of the field. However, this variety also implies a restriction, because it complicates the direct comparison of the results and the elaboration of the most accurate quantitative conclusions.

Taken together, the findings suggest that microgravity has a significant influence on the physiology of red blood cells, impacting not only their quantity but also their capacity and functional quality. Thus, osmotic fragility appears as a useful indicator to analyze the integrity of erythrocytes, although its investigation still needs further investigation.

Finally, this research helps to unite the existing evidence and also to point out the gaps present in the literature, underlining the urgency of further investigating this phenomenon through more consistent studies and with comparable methodologies, as they facilitate a better understanding of the mechanisms involved and their effect on human health in microgravity situations.

VII. Recommendations for Future Studies

Based on the results in the literature, the following recommendations are proposed to guide future research on the impacts of microgravity on red blood cells. For example, create and use standardized techniques to assess osmotic fragility, incorporating the use of OFT curves and parameters such as MCF/H50 (Mean Corpuscular Fragility-H50), which will allow an objective comparison between the multiple investigations. Evaluate research that includes various variables, such as erythrocyte deformability, hemolysis, oxidative stress, and membrane changes, in order to understand the most comprehensive form of the mechanisms involved. Promote the standardization of experimental protocols and reporting criteria, as well as the use of larger and more varied samples, to support the validity and replicability of the results. To carry out experimental and longitudinal research that facilitates the evaluation of variations in erythrocytes over time in order to deepen the evolution of hemolysis and cellular vulnerability in microgravity environments. Finally, increase the number of studies in conditions of true microgravity, as well as carry out direct comparisons with simulated models, in this sense to be able to corroborate the validity of the results obtained in the laboratory.

References

[1] Almeida, C. M. Juran, K. M. Brown, and E. Mistry, "Effects of Mechanical Unloading on Hematopoiesis and Post-flight Red Blood Cell Lifecycle," NASA Technical Reports Server, 2024. [Online]. Available:

Https://ntrs.nasa.gov/citations/20240002361https://ntrs.nasa.gov/citations/20240002361

[2] N. Kashirina, a. G. Brzhozovskiy, W. Sun, L. K. Pastushkova, O. V. Popova, V. B. Rusanov Et Al., "Proteomic Characterization of Dry Blood Spots of Healthy Women during Simulation the Microgravity Effects Using Dry Immersion," Frontiers in Physiology, Vol. 12, P. 753291, 2022, Doi: 10.3389/fphys.2021.753291.

[3] A. M. Abolyazed, W. a. Elbehery, H. M. Elsayed Et Al., "Effects of Simulated Microgravity on Human Hematological and Anemia-related Biomarkers: a Systematic Review and Meta-analysis," Bulletin of the National Research Centre, Vol. 50, Art. No. 19, 2026, DOI: 10.1186/S42269-026-01407-W.

[4] Buchheim, M. Feuerecker, M. Balsamo, M. Vukich, M. Van Walleghem, K. Tabury Et Al., "Monitoring Functional Immune Responses with a Cytokine Release Assay: ISS Flight Hardware Design and Experimental Protocol for Whole Blood Cultures Executed under Microgravity Conditions," Frontiers in Physiology, Vol. 14, P. 1322852, 2024, Doi: 10.3389/fphys.2023.1322852.

[5] Hinge, J. Dixit, P. Vidyasagar and G. Kulkarni, "Comparative Analysis of Red Blood Cell under Real and Simulated Microgravity Conditions," bioRxiv, 2024, Doi: 10.1101/2024.12.26.630384.

[6] Ramos, E. Cantó, a. Mercado, V. Romero-Alva, N. I. Vargas-Cuentas, and a. Roman-Gonzalez, "Student Experiences with a Clinostat for Microgravity Simulation," IEEE Aerospace and Electronic Systems Magazine, Vol. 40, No. 5, Pp. 76–80, 2025, Doi: 10.1109/MAES.2025.3546146.

[7] Lansiaux, N. Jain, S. Y. Chodnekar, a. Siddiq, M. Ibrahim, M. Yèche, and I. Kantane, "Understanding the Complexities of Space Anaemia in Extended Space Missions," Frontiers in Physiology, Vol. 15, P. 1321468, 2024, Doi: 10.3389/fphys.2024.1321468.

[8] Muralidhara, J. K. Pachiyappan, a. Marimuthu, G. Y. Vassilieva, B. S. S. Nidamanuri, and G. Kuppusamy, "Microgravity-induced Anaemia: Insights, Mechanisms, and Inducing Factors," Microgravity Science and Technology, Vol. 38, No. 1, Art. No. 3, 2025, DOI: 10.1007/S12217-025-10226-Z.

[9] E. Igbineweka and J. J. Van Loon, "Gene-environmental Influence of Space and Microgravity on Red Blood Cells with Sickle Cell Disease," Npj Genomic Medicine, Vol. 9, No. 1, Art. No. 44, 2024, DOI: 10.1038/S41525-024-00427-7.

[10] Agramont, M. Carrasco, B. Flores, D. Quenta, and G. Chávez, "Osmotic Fragility and Its Relationship with Space Anemia," Revista Ciencia Y Cultura, Vol. 27, No. 51, Pp. 91–106, 2023.

[11] Manis, a. Manca, a. Murgia, G. Uras, P. Caboni, T. Congiu Et Al., "Understanding the Behaviour of Human Cell Types under Simulated Microgravity Conditions: the Case of Erythrocytes," International Journal of Molecular Sciences, Vol. 23, No. 12, P. 6876, 2022, Doi: 10.3390/ijms23126876

[12] Murali and R. R. Sarkar, "Impact of Gravitational Forces on Red Blood Cell Dynamics in Biofluid Suspension," Life Sciences in Space Research, Vol. 47, Pp. 197–210, 2025, Doi: 10.1016/j.lssr.2025.06.012.

[13] Manis, "Study of Metabolic Changes in Human Erythrocytes and Plasma under Simulated Microgravity Conditions," 2023.

[14] Trudel, N. Shahin, T. Ramsay, O. Laneuville, and H. Louati, "Hemolysis Contributes to Anemia during Long-duration Space Flight," Nature Medicine, Vol. 28, No. 1, Pp. 59–62, 2022, Doi: 10.1038/s41591-021-01637-7.

[15] Bogdanova, L. Kaestner, G. Simionato, a. Wickrema, and a. Makhro, "Heterogeneity of Red Blood Cells: Causes and Consequences," Frontiers in Physiology, Vol. 11, P. 392, 2020, Doi: 10.3389/fphys.2020.00392.

[16] Trudel, N. Shahin, T. Ramsay, O. Laneuville, and H. Louati, "Hemolysis Contributes to Anemia during Long-duration Space Flight," Aerospace Medicine and Human Performance, Vol. 94, No. 6, Pp. 444–450, 2023. Doi: 10.3357/AMHP.6190.2023

[17] Zheng, J. Charvat, S. R. Zwart, S. K. Mehta, B. E. Crucian, S. M. Smith Et Al., "Time-resolved Molecular Measurements Reveal Changes in Astronauts during Spaceflight," Frontiers in Physiology, Vol. 14, P. 1219221, 2023, Doi: 10.3389/fphys.2023.1219221.

[18] S. Olver, "Erythrocyte Structure and Function," in Schalm's Veterinary Hematology, 7th Ed., D. J. Weiss, K. J. Wardrop, and J. Schaefer, Eds. Hoboken, NJ, USA: Wiley-Blackwell, 2022, Pp. 158–165. Doi: 10.1002/9781119500537.ch20 10.1002/9781119500537.ch20

[19] Obeagu, M. C. Igwe, and G. U. Obeagu, "Oxidative Stress's Impact on Red Blood Cells," Medicine, Vol. 103, No. 9, P. E37360, 2024, Doi: 10.1097/MD.0000000000037360.

[20] N. Chatzinikolaou Et Al., "Erythrocyte Metabolism," Acta Physiologica, Vol. 240, No. 3, P. E14081, 2024, Doi: 10.1111/apha.14081.

[21] Hicks, M. Olson, C. Mitchell, C. M. Juran, and a. M. Paul, "the Impact of Microgravity on Immunological States," ImmunoHorizons, Vol. 7, No. 10, Pp. 670–682, 2023, Doi: 10.4049/immunohorizons.2200063.

[22] Azzi, "Oxidative Stress: What Is it?," Antioxidants, Vol. 11, No. 8, P. 1431, 2022, Doi: 10.3390/antiox11081431.

[23] Von Petersdorff-Campen and M. S. Daners, "Hemolysis Testing in Vitro: a Review," ASAIO Journal, Vol. 68, No. 1, Pp. 3–13, 2022, Doi: 10.1097/MAT.0000000000001454.

[24] Jacob Et Al., "Erythrocyte Deformability Correlates with Systemic Inflammation," Blood Cells, Molecules, and Diseases, Vol. 109, P. 102881, 2024, Doi: 10.1016/j.bcmd.2024.102881.

[25] J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. a. Akl, S. E. Brennan, Et Al., "the PRISMA 2020 Statement: an Updated Guideline for Reporting Systematic Reviews," Revista Panamericana De Salud Pública, Vol. 46, P. E112, 2022, Doi: 10.26633/RPSP.2022.112.

[26] Bognár, B. Szentes, and P. Benedek, "Development of the PRISM Risk Assessment Method," Risks, Vol. 10, No. 11, P. 213, 2022, Doi: 10.3390/risks10110213.

[27] Mishra and M. P. Mishra, "PRISMA for Review of Management Literature," 2023, Doi: 10.1108/S2754-586520230000002007.

[28] Tan, Z. Zhou, H. Zheng, Y. Li, H. Wang, Q. Yang, H. Tian, H. Chen, J. Xie, Z. Li, and Y. Chen, "Emerging Themes and Future Directions in Space Radiation Health Research: a Bibliometric Exploration from 2013 to 2022," Radiation and Environmental Biophysics, Vol. 64, No. 2, Pp. 211–227, 2025, Doi: 10.1007/s00411-025-01115-5

[29] J. Dino, C. M. Villafuerte, J. Pimentel, J. Sayat, J. C. Vital, D. Hernandez, L. Thiamwong, V. Decker, M. Shattell, L. E. Cruz, M. Leoncio, J. R. Reyes, J. David, and J. a. Manalili, "Aerospace Health: a Systematic Review and Current State of Science," Frontiers in Space Technologies, Vol. 7, Art. No. 1736627, 2026, DOI: 10.3389/FRSPT.2026.1736627.

[30] Li, C. Deacon, and M. R. Keezer, "the Performance of Bibliometric Analyses in the Health Sciences," Current Medical Research and Opinion, Vol. 40, No. 1, Pp. 1–11, 2024, Doi: 10.1080/03007995.2023.2281503.

[31] Passas, "Bibliometric Analysis: the Main Steps," Encyclopedia, Vol. 4, No. 2, Pp. 1014–1025, 2024, Doi: 10.3390/encyclopedia4020065.

[32] Gu, C. Gao, and L. Wang, "the Evolution of Artificial Intelligence in Biomedicine: Bibliometric Analysis," JMIR AI, Vol. 2, P. E45770, 2023, Doi: 10.2196/45770.

[33] Kumar, "Bibliometric Analysis: Comprehensive Insights into Tools, Techniques, Applications, and Solutions for Research Excellence," Scientific and Educational Management Studies, Vol. 3, No. 1, Pp. 1–16, 2025, Doi: 10.31181/sems31202535k.

[34] Fu, H. Xu, Y. Lai, X. Sun, Z. Zhu, H. Zang, and Y. Wu, "a VOSviewer-Based Bibliometric Analysis of Prescription Refills," Frontiers in Medicine, Vol. 9, Art. No. 856420, 2022, DOI: 10.3389/FMED.2022.856420.

[35] Guleria and G. Kaur, "Bibliometric Analysis of Ecopreneurship," Library Hi Tech, Vol. 39, No. 4, Pp. 1001–1024, 2021, Doi: 10.1108/LHT-09-2020-0218.

[36] Rossello Et Al., "EURObservational Research Programme," European Heart Journal, Vol. 8, No. 8, Pp. 804–811, 2022, Doi: 10.1093/ehjqcco/qcac041.

[37] Wondimagegn Et Al., "Bibliometric Analysis of Author Distribution," BMJ Global Health, Vol. 8, No. 6, P. E011656, 2023, Doi: 10.1136/bmjgh-2022-011656.

[38] P. Costa Et Al., "Integrating Bibliometrics and Qualitative Content Analysis," 2023, Doi: 10.1109/ACIT58888.2023.10453680.

[39] N. Lyhne, J. Thisted, and M. Bjerrum, "Qualitative Content Analysis," Quality & Quantity, Vol. 59, Pp. 5329–5349, 2025, Doi: 10.1007/s11135-025-02220-9.

[40] Ghasemi, P. Mirmiran, K. Kashfi, and Z. Bahadoran, "Scientific Publishing in Biomedicine," International Journal of Endocrinology and Metabolism, Vol. 21, No. 1, 2022, Doi: 10.5812/ijem-131812.

[41] A. Gerda Et Al., "Comparative Analysis of Erythrocyte Osmotic Fragility," Journal of Evolutionary Biochemistry and Physiology, Vol. 60, No. 4, Pp. 1363–1384, 2024, Doi: 10.1134/S0022093024040094.

[42] J. Kim, S. Park and S. Kim, "Comparison of Clinostat Control Strategies to Achieve Simulated Microgravity with Uniform Gravity Vector Distribution," Npj Microgravity, Vol. 12, No. 1, Art. No. 21, 2026, DOI: 10.1038/S41526-026-00570-8.

[43] Fais, F. Ghiani, D. Dessì, M. Casula, G. Perra, E. Torchia, N. Lai, G. Cao and a. Concas, "Superfood Potential of Chlorella Vulgaris: Productivity and Antioxidant Boost under Simulated Moon and Microgravity Conditions," Npj Microgravity, Vol. 12, No. 1, Art. No. 8, 2026, DOI: 10.1038/S41526-025-00550-4.

Artificial Intelligence-Assisted Community Needs Assessment and Extension Planning: Evidence from Wesleyan University-Philippines Partner Communities

Eufemia Ayro[1], Karl Leugim Bernarte[2*], Hazel May Babiera[3], Evangeline Agpoon[4], Jhon Carlo Villa[5], Maureen Bondoc[6], Jennyfer Villalon[7], Jose Arsenio Adriano[8], Christian Navarro[9]
Graduate School, Wesleyan University-Philippines, Cabanatuan City, Philippines[1, 2, 3, 4, 5, 6, 7]
WESCOP, Wesleyan University-Philippines, Cabanatuan City, Philippines[1, 8, 9]

***Abstract*—This study conducted an Artificial Intelligence-assisted community needs assessment of partner-community respondents of Wesleyan University-Philippines, with emphasis on the Tricycle Operators and Drivers Association. Using a descriptive-quantitative design, five objectives were addressed: to describe the socioeconomic profile of the respondents; to determine livelihood and income conditions; to identify health, educational, and public-service needs; to rank priority community needs; and to develop a transparent, data-driven Artificial Intelligence-assisted framework. A cleaned dataset of 151 respondents was analyzed, of whom 46 were formally affiliated with tricycle operator and driver associations. Frequency, percentage, mean, and a proportion-based priority score summarized the data, while an Artificial Intelligence-assisted workflow supported response coding, respondent clustering, pattern detection, and need-to-intervention matching. Results showed that the transport subgroup consisted entirely of male drivers with lower average income, stronger income-expenditure pressure, greater reliance on borrowing and relatives, and recurring health concerns, including respiratory illness, hypertension, dental problems, and diabetes. Educational needs centered on books and learning materials. Ranked priorities were transport-oriented support, public-service linkage, health and wellness, educational assistance, livelihood and financial stability, skills training, and environmental support. The study concludes that pairing descriptive statistics with a reproducible Artificial Intelligence-assisted procedure produces more targeted, equitable, and responsive extension planning, and offers a practical template for other higher-education institutions.**

***Keywords*—Artificial Intelligence; Community Needs Assessment; data-driven extension program; tricycle drivers; needs prioritization; decision support**

I. Introduction

Community needs assessment is a critical process for identifying the social, economic, health, livelihood, and service-related concerns of a target population. In university extension work, it provides an empirical basis for designing programs that respond to the actual conditions of partner communities rather than to assumptions or generalized interventions. Prior studies have shown the value of structured surveys for determining industry and community priorities, building extension programs, and identifying future research and service needs [1]–[6]. These studies indicate that systematic data collection can help institutions align their programs with the needs of specific beneficiary groups.

At Wesleyan University-Philippines, the Tricycle Operators and Drivers Association represents an important community sector. Tricycle drivers contribute directly to local mobility, campus access, and everyday transportation. Like other transport-dependent and underserved groups, they may face income instability, high household expenses, limited access to public services, unmet health needs, and gaps in skills development. Transportation research has emphasized that mobility systems are closely connected to social equity, public participation, and access to opportunities [7]–[11]. Assessing the needs of these members is therefore not only a question of livelihood analysis but also one of inclusive community planning and equitable service delivery.

Public participation is also essential to identifying community priorities. Arnstein's framework on citizen participation stresses the importance of moving beyond token consultation toward meaningful involvement in decision-making [12]. Later work similarly argues that effective planning requires participatory approaches that let affected groups express their needs and lived experiences [13]–[17]. For transport workers, participatory needs assessment can help ensure that university extension initiatives address genuine concerns such as livelihood support, financial literacy, health services, skills training, and access to institutional assistance.

Traditional needs assessment, however, often produces large amounts of survey data that require careful interpretation. Artificial Intelligence offers a way to improve this analysis by assisting in pattern detection, classification, prioritization, and decision support. Applied responsibly, it can help identify which needs are most urgent, which respondent groups are most vulnerable, and which interventions are most appropriate for a given community profile. This direction is consistent with survey-based studies that used structured instruments and quantitative analysis to evaluate community or industry needs [5], [18], [19], and with a growing body of work that applies machine learning to community and poverty data [26]–[31].

*Corresponding author

Equity-oriented planning literature further supports examining communities that may be disadvantaged, underrepresented, or underserved in formal decision-making [20]–[23], [9], [11]. Studies on transportation equity and public engagement show that vulnerable communities are often affected by planning decisions yet have limited influence over the programs intended to serve them [24], [9]–[11], [25]. An Artificial Intelligence-assisted needs assessment can provide a more systematic way to prioritize interventions and ground extension programs in evidence, equity, and participation.

This study analyzed all 151 respondents in the dataset, while giving particular emphasis to the 46 respondents who were formally affiliated with tricycle operator and driver associations. The emphasis is justified because these members form a coherent transport-sector subgroup that shares an occupation, an income structure, and a set of road-related exposures; retaining the full sample allows their conditions to be benchmarked against the broader partner community. Accordingly, the study addressed five objectives: 1) to describe the socioeconomic profile of the respondents; 2) to determine their livelihood and income conditions; 3) to identify their health, educational, and public-service needs; 4) to rank their priority community needs; and 5) to develop a data-driven, Artificial Intelligence-assisted community extension framework.

The contribution of the study is threefold. First, it documents a transparent and reproducible Artificial Intelligence-assisted workflow—covering response coding, clustering, pattern detection, proportion-based prioritization, and need-to-intervention matching—rather than treating Artificial Intelligence as a black box. Second, it shows how this workflow adds value beyond conventional frequency-and-percentage reporting by isolating a vulnerable subgroup and linking each surfaced need to a concrete extension response. Third, it provides a deployable framework that a university can reuse for planning, implementation, monitoring, and evaluation of extension programs for transport and other partner communities.

II. Related Work

A. Community Needs Assessment and University Extension

Structured surveys have long been used to characterize the needs of specific industries and communities and to translate those needs into extension programs [1]–[6]. This literature established that descriptive instruments, careful coding, and ranking can yield actionable priorities, and that ongoing evaluation is needed to keep programs aligned with beneficiaries [3]. Community-engaged health-needs assessments similarly combine closed and open-ended items with descriptive and thematic analysis to direct concrete public-health initiatives [28]. The present study builds on this tradition but extends it to a transport-sector partner community of a university.

B. Transportation Communities, Equity, and Informal Transport Workers

Transportation scholarship frames mobility as a matter of social equity and meaningful participation, noting that underserved communities are frequently shaped by decisions they have little power to influence [7]–[11], [24], [25]. Within this literature, informal transport workers are a distinctly vulnerable group. Drivers of tricycles, pedicabs, and similar vehicles face poverty risks linked to low educational attainment, prolonged participation in the informal economy, and limited social protection, as recent machine-learning and econometric analyses of informal transportation in the Philippines demonstrate [26], [27]. These conditions motivate a needs assessment that treats tricycle operators and drivers as a priority sector rather than as an undifferentiated part of the population.

C. Artificial Intelligence Needs Assessment and Decision Support

A rapidly expanding body of work applies machine learning to survey and administrative data for social purposes. Studies have used supervised and tree-based models to predict healthcare needs and access barriers from community surveys [28], and to predict and target poverty across localities, often finding ensemble methods such as random forests effective [27], [29], [30]. For unstructured responses, topic modeling, embedding-based clustering, and few-shot classification with language models now support faster and more consistent coding of open-ended items, while still requiring human interpretation of the resulting clusters [31]. Collectively, this literature positions Artificial Intelligence as a decision-support layer that augments, rather than replaces, descriptive analysis and expert judgment.

D. Research Gap and Positioning of the Study

Two gaps remain. First, machine-learning poverty and needs studies typically operate at a regional or national scale and rarely target a single university partner community, such as a campus-serving tricycle association. Second, many "Artificial Intelligence-assisted" community studies describe the technology only in general terms, without specifying the workflow, inputs, and outputs. This study addresses both gaps by focusing on a defined transport-sector partner community and by documenting an explicit, reproducible Artificial Intelligence-assisted procedure that maps surfaced needs to specific extension responses. In doing so, it differs from conventional survey-based assessment, which usually stops at frequency, percentage, mean, and ranking, by adding respondent grouping, pattern detection, and need-to-intervention matching that can be carried into program monitoring.

III. Methodology

A. Research Design

This study employed a descriptive-quantitative research design integrated with Artificial Intelligence-assisted data analysis to assess and prioritize the needs of members of the Tricycle Operators and Drivers Association of Wesleyan University-Philippines and of the broader partner community. The descriptive component determined the demographic, socioeconomic, livelihood, health, educational, skills-development, and public-service profile of the respondents. The Artificial Intelligence-assisted component supported response coding, clustering, pattern recognition, and prioritization of needs.

The design is appropriate because community needs assessment studies commonly use structured survey data to identify priority concerns and guide extension programs [1]–[6]. Here, Artificial Intelligence served as a decision-support tool

that helped organize and interpret patterns in the dataset. The overall methodological flow is shown in Table I.

TABLE I. RESEARCH DESIGN MATRIX

Component	Description	Purpose	Expected Output
Research Design	Descriptive-quantitative with Artificial Intelligence-assisted analysis	To describe and analyze the needs of the respondents.	Community needs profile
Data Source	Community Needs Assessment Survey	To gather demographic, socioeconomic, health, livelihood, and service-related data.	Encoded dataset
Statistical Analysis	Frequency, percentage, mean, and proportion-based priority score	To identify dominant respondent characteristics and priority needs.	Summary tables and ranked needs
Artificial Intelligence-Assisted Analysis	Coding, clustering, pattern recognition, and prioritization	To group respondents and determine high-priority intervention areas.	Needs-priority model
Final Output	Data-driven extension program framework	To guide university extension planning.	Proposed extension program

The overall methodological flow of the study is presented in Fig. 1.

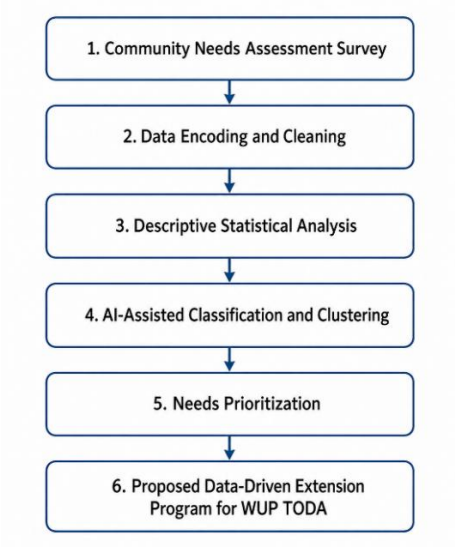


Fig. 1. Methodological flow of the AI-assisted community needs assessment.

B. Research Locale

The study was conducted at Wesleyan University-Philippines, with emphasis on tricycle operators and drivers working within and around the university premises. The dataset included respondents associated with WUP TODA, WUP Main Gate TODA, TODA WUP, and other tricycle associations near the university. The locale was selected because these members provide daily transportation to students, employees, and nearby residents, which makes them a relevant partner community for university-based extension services.

C. Respondents of the Study

The study used secondary data from an existing Community Needs Assessment, so no new sampling was performed by the researchers during analysis. The dataset was treated as a complete enumeration of the available respondents: all 151 records that passed data screening were analyzed rather than a probability sample drawn from them. Within this total enumeration, a purposive emphasis was placed on the 46 respondents who were formally affiliated with tricycle operator and driver associations, because they constitute the priority transport sector for the university's extension work. The combined approach is therefore a complete enumeration of an available, non-probability dataset with a purposively defined subgroup. This choice is consistent with the descriptive aim of the study, but it limits statistical generalization beyond the participating community, a point revisited in the limitations.

The respondents were members of the partner community of Wesleyan University-Philippines, including tricycle drivers, operators, and household representatives recorded in the Community Needs Assessment dataset. They were described according to demographic and socioeconomic variables such as age, sex, civil status, educational attainment, occupation, household size, monthly income, expenses, skills, health condition, access to services, and identified priority needs (Table II).

TABLE II. PROFILE VARIABLES OF THE RESPONDENTS

Variable Category	Specific Variables	Purpose of Analysis
Demographic Profile	Age, sex, civil status, household size	Described the general characteristics of partner community respondents.
Educational Profile	Highest educational attainment and school-age dependents	Identified education-related needs and possible learning support programs.
Economic Profile	Occupation, monthly income, and household expenses	Assessed income stability and financial vulnerability.
Livelihood Profile	Main source of income, additional livelihood, and skills	Determined livelihood and skills-training needs.
Health Profile	Health concerns and access to health services	Identified possible medical, wellness, or health-extension needs.
Public Service Access	Access to government or community services	Determined gaps in service delivery.
Priority Needs	Stated household and community needs	Ranked intervention areas for extension planning.

D. Research Instrument

The primary instrument was the Community Needs Assessment Survey Form. It contained items on respondent profile, socioeconomic condition, livelihood status, educational background, health concerns, available skills, access to services, environmental concerns, and priority community needs. Scaled or selection-type responses were interpreted using proportion-based scoring and ranking, consistent with structured measurement in survey research [18]. The survey results also

served as the input dataset for the Artificial Intelligence-assisted analysis.

E. Data Gathering Procedure

The study used the existing Community Needs Assessment data collected from respondents of Wesleyan University-Philippines. The data were reviewed, encoded, cleaned, and organized prior to statistical and Artificial Intelligence-assisted analysis. The procedure followed seven steps: responses were collected and organized; reviewed for completeness, consistency, and relevance; checked and corrected for duplicate, unclear, or inconsistent entries; coded into analyzable categories for categorical and text-based items; prepared for descriptive and Artificial Intelligence-assisted processing; summarized into tables, patterns, clusters, and rankings; and finally translated into a proposed data-driven extension program.

F. Data Preparation, Coding, and Reliability

Before analysis, the dataset was preprocessed to improve accuracy and consistency. Missing values, inconsistent entries, duplicate responses, and unclear categories were checked. Categorical responses were standardized, and open-ended responses were grouped into common themes (Table III). For the Artificial Intelligence-assisted analysis, categorical variables were converted into coded values, and multiple-response items such as skills, services needed, and priority concerns were transformed into binary indicators.

To strengthen the thematic grouping of open-ended responses, a documented coding protocol was applied. A draft codebook of recurring themes was first prepared from an initial read of the responses. A language-model assistant proposed candidate theme labels for each response, and two researchers independently reviewed and assigned the final codes against the codebook. Disagreements were resolved by discussion and consensus, and a sample of coded responses was re-checked for consistency. In this way, the Artificial Intelligence output was used as a suggestion that researchers verified, not as a final categorization, which preserves human control over interpretation.

TABLE III. DATA PREPROCESSING AND CODING PLAN

Data Type	*Example Variables*	*Preprocessing Procedure*	*Output*
Numeric Data	Age, income, expenses, household size	Missing values and outliers were checked.	Clean numeric dataset
Categorical Data	Sex, civil status, occupation, educational attainment	Responses were standardized and coded.	Coded variables
Text-Based Data	Skills, stated needs, livelihood concerns	Similar responses were grouped into themes.	Thematic categories
Multiple-Response Data	Priority needs, services needed, training interests	Responses were converted into binary or frequency indicators.	Needs matrix
Incomplete Entries	Blank or inconsistent responses	Entries were validated, corrected, or excluded when necessary.	Final cleaned dataset

G. Statistical Treatment of Data

Descriptive statistics summarized the profile and needs of the respondents. Frequencies and percentages were used for categorical variables, and the mean was used for continuous variables such as age, income, and expenditure. The percentage for category i was computed as $\%_i = (f_i / n) \times 100$, where f_i is the frequency of category i and n is the number of valid responses. The arithmetic mean was computed as $\bar{x} = (\Sigma x_i) / n$.

For multiple-response and binary-coded needs, a proportion-based priority score (endorsement rate) was used. For need j, the score was $ER_j = (\Sigma_i b_{ij}) / n$, where $b_{ij} = 1$ if respondent i endorsed need j and 0 otherwise, and n is the relevant denominator (151 for all respondents and 46 for the transport subgroup). The score ranges from 0 to 1 and equals the proportion of respondents who indicated the need; needs were then ranked in descending order of ER_j. This formulation is reported transparently because the values in Table IV are endorsement rates, and the column is labeled accordingly. Computational tools were used to organize and process the dataset, consistent with survey-based studies that applied statistical software [19].

TABLE IV. STATISTICAL TREATMENT OF DATA

Research Objective	*Data Required*	*Statistical Tool*	*Interpretation*
Describe the profile of respondents	Demographic and socioeconomic variables	Frequency and percentage	Identified dominant respondent characteristics.
Determine livelihood and income conditions	Income, occupation, expenses, livelihood source	Frequency, percentage, and mean	Showed economic status and livelihood vulnerability.
Identify health, education, and service needs	Health concerns, educational needs, public-service access	Frequency and ranking	Determined common social-service needs.
Rank priority community needs	Stated needs and intervention preferences	Endorsement rate (proportion) and rank	Identified the most urgent areas for extension.
Develop an Artificial Intelligence-assisted framework	Ranked needs, cleaned and coded dataset, profile and livelihood indicators	Coding, normalization, pattern identification, proportion-based prioritization, framework mapping	Produced a framework identifying priority intervention areas.

H. Artificial Intelligence-Assisted Data Analysis

Artificial Intelligence was integrated to support deeper interpretation of the dataset, focusing on coding, clustering, pattern recognition, and needs prioritization. To make the procedure transparent and reproducible, the specific tools and workflow are described here. A general-purpose large language model assistant was used for assisted thematic coding of open-ended responses, and a Python environment with the pandas and scikit-learn libraries was used for the quantitative steps.

The workflow proceeded in seven steps: 1) The cleaned and coded dataset was ingested. 2) Open-ended responses were thematically coded with language-model assistance and researcher review, as described in Section III G. 3) Categorical

and multiple-response items were converted into numeric and binary features. 4) Respondents were grouped using k-means clustering on standardized socioeconomic features, while text needs were grouped using term-frequency representations and similarity-based clustering. 5) Patterns were detected through cross-tabulation and association checks, for example, between income level and livelihood needs, household size and financial difficulty, or occupational exposure and health concerns. 6) Needs were prioritized using the endorsement-rate score defined above. 7) Each high-priority need was matched to a candidate extension response through a rule-based mapping that researchers reviewed. The input variables and expected outputs of each procedure are summarized in Table V, and the prioritization logic is illustrated in Fig. 2.

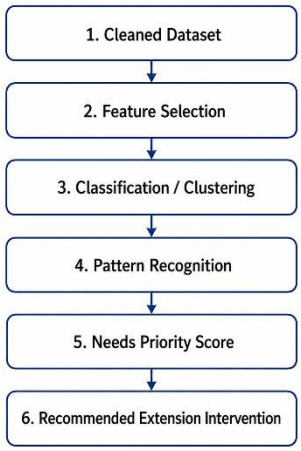


Fig. 2. AI-assisted needs prioritization framework.

TABLE V. AI-ASSISTED ANALYSIS PLAN

AI Procedure	*Purpose*	*Possible Input Variables*	*Expected Output*
Classification / Coding	Categorized respondents and responses by dominant need	Income, occupation, skills, health and education needs	Respondent and response need categories
Clustering	Grouped respondents with similar conditions	Socioeconomic and household variables	Clusters of similar respondents
Pattern Recognition	Detected relationships among variables	Income, expenses, household size, livelihood, needs	Identified trends and associations
Priority Modeling	Determined urgent intervention areas	Endorsement rate of needs, respondent profile	Ranked community priorities
Decision Support	Guided extension program design	Artificial Intelligence and statistical results	Proposed intervention framework

The Artificial Intelligence-assisted analysis was not a replacement for the researcher's interpretation. It functioned as a decision-support mechanism that revealed patterns and priority areas. Its value beyond conventional descriptive statistics is summarized as follows: whereas frequency, percentage, mean, and ranking describe the dataset one variable at a time, the assisted workflow additionally grouped respondents into comparable clusters, detected co-occurring needs across variables, isolated the more vulnerable transport subgroup, and linked each surfaced need to a specific intervention. These capabilities turn a set of summary tables into a reusable decision-support layer for planning, implementation, and monitoring.

I. Needs Prioritization Procedure

Priority needs were determined using both the endorsement-rate ranking and the Artificial Intelligence-assisted analysis. Needs that appeared frequently across respondents were identified and ranked, then examined in relation to profile variables such as income, household expenses, occupation, health status, household size, and available skills. A prioritization matrix connected each identified need with a possible university extension response (Table VI).

TABLE VI. NEEDS PRIORITIZATION MATRIX

Priority Area	*Basis for Ranking*	*Possible Extension Response*
Livelihood and Income Support	Low income, high expenses, and unstable livelihood	Livelihood training and income-generating projects
Skills Development	Existing skills and training interests	Technical, vocational, or entrepreneurial training
Health and Wellness	Reported health concerns or limited access to health services	Medical mission, health education, and wellness programs
Financial Literacy	Income-expense gap and household financial difficulty	Budgeting seminar, savings orientation, and microenterprise training
Educational Support	Education-related household needs	Scholarship orientation, tutorial support, and learning assistance
Public Service Access	Limited access to government or community services	Linkage with local government units and social-service agencies

J. Validation of Results

The statistical and Artificial Intelligence-assisted results were reviewed to ensure that the generated priorities were reasonable, interpretable, and aligned with the actual responses. Validation compared the assisted outputs with the descriptive statistical results and the stated needs of the respondents. The interpretation also considered the importance of meaningful community participation in planning, because participatory approaches strengthen the relevance of community-based programs by allowing affected groups to express their actual needs and priorities [13]–[17], [24], [25].

K. Ethical Considerations

The study observed ethical standards in handling community data. Respondent information was treated confidentially, and personal identifiers were removed before analysis. The dataset was used only for academic and extension-planning purposes. Because the study involved Artificial Intelligence-assisted analysis, the generated outputs were reviewed carefully before interpretation. The researchers ensured that findings were presented in a respectful, community-centered, and equity-oriented manner, and that results were used to identify appropriate support programs rather than to stigmatize respondents.

IV. Results and Discussions

A. Data Screening and Scope of Analysis

The Community Needs Assessment dataset contained 151 respondents, and the analysis covered all of them. Because the study prioritized tricycle operators and drivers of Wesleyan University-Philippines, a transport-focused comparison was also made. Of the 151 respondents, 46 (30.5%) were formally affiliated with tricycle associations, including WUP TODA, WUP Main Gate TODA, TODA WUP, Mabini Ext TODA, and a senior-citizen WUP TODA entry; among them, 36 (23.8%) were directly linked to WUP-related groups. Multiple-response items were treated as selected when the corresponding cell contained an entry, so percentages in some tables may exceed 100%.

This emphasis on the transport subgroup is methodologically defensible: the 46 members share an occupation, an income structure, and a set of road-related exposures, which makes them a coherent unit for targeted planning, while the full sample provides a comparison baseline. The framing as a transport-focused study therefore reflects the analytical priority rather than an overstatement of the subgroup's share, and the title has been revised to make the broader respondent base explicit.

B. Profile of the Respondents

Frequency and percentage were used to identify dominant respondent characteristics (Table VII). The general respondent group was predominantly male, with 110 male respondents (72.8%). This pattern was stronger in the transport subgroup, where all 46 members (100.0%) were male, indicating that transport-related livelihood in the dataset was male-dominated.

By age, the largest group among all respondents was 40–49 years (63; 41.7%), and the same pattern held for the transport subgroup (14; 30.4%). The mean age was 45.78 years overall and 46.91 years for the transport subgroup, indicating a middle-adult population for whom household, livelihood, health, and educational responsibilities are typically high. For occupation, 80 respondents (53.0%) were in transport or driver-related work; among the transport subgroup, 45 of 46 (97.8%) were transport workers, confirming a clearly defined transport-sector community within the larger dataset.

TABLE VII. Profile of all Respondents with TODA Emphasis

Profile Variable	Category	All f	All %	TODA f	TODA %
Sex	Male	110	72.8	46	100.0
	Female	41	27.2	0	0.0
Age Group	Below 30	5	3.3	2	4.3
	30–39	37	24.5	13	28.3
	40–49	63	41.7	14	30.4
	50–59	31	20.5	7	15.2
	60 and above	14	9.3	9	19.6
	Not reported	1	0.7	1	2.2
Occupation	Transport/driver-related	80	53.0	45	97.8
	Other / not reported	71	47.0	1	2.2

C. Livelihood and Income Conditions

Frequency, percentage, and mean were used to show economic status and livelihood vulnerability (Table VIII). Valid monthly income data were available for 100 respondents. The mean monthly income of all respondents was ₱8,945.20 and the median was ₱9,250.00. For the transport subgroup, valid income data were available for 41 respondents, with a lower mean of ₱7,682.93 and a median of ₱9,000.00.

Income was concentrated at the lower ranges. In the full sample, 52 respondents (34.4%) earned below ₱10,000 per month; among the transport subgroup, the share rose to 26 of 46 (56.5%). This greater concentration of low income is consistent with machine-learning and econometric findings that informal transport workers in the Philippines face elevated poverty risk tied to low educational attainment and prolonged informal-sector participation [26], [27].

TABLE VIII. Monthly Income Distribution

Monthly Income Range	All f	All %	TODA f	TODA %
Below ₱5,000	24	15.9	14	30.4
₱5,000–₱9,999	28	18.5	12	26.1
₱10,000–₱14,999	31	20.5	12	26.1
₱15,000–₱19,999	16	10.6	3	6.5
₱20,000 and above	1	0.7	0	0.0
Not reported / non-numeric	51	33.8	5	10.9

Household expenditure patterns reinforced this picture (Table IX). Food was the most frequently reported expense overall, followed by electricity or gas, education, water, and cooking fuel. Among the transport subgroup, food, electricity or gas, and water were equally common, each reported by 18 respondents (39.1%). Valid total monthly expenditure was available for 48 respondents, with a mean of ₱11,044.17 and a median of ₱9,275.00; for the transport subgroup (12 respondents), the mean was ₱8,438.33 and the median ₱8,300.00.

TABLE IX. Common Household Expenditure Categories

Expenditure Category	All f	All %	TODA f	TODA %
Food	63	41.7	18	39.1
Electricity / Gas	57	37.7	18	39.1
Education	51	33.8	12	26.1
Water	46	30.5	18	39.1
Cooking Fuel	40	26.5	13	28.3
Soap	27	17.9	13	28.3
Clothing	26	17.2	14	30.4

Among respondents with both valid income and expenditure data, 20 of 39 (51.3%) had expenses exceeding income; among the transport subgroup, 8 of 12 (66.7%) did. This stronger income-expenditure pressure helps explain the coping strategies reported (Table X). The most common strategy overall was

borrowing money (73; 48.3%), whereas in the transport subgroup the most common was asking relatives for help (25; 54.3%).

These findings indicate heavy dependence on informal support networks. While family assistance can cushion financial difficulty, reliance on borrowing and on relatives also signals economic instability, echoing accounts of the financial precarity of Filipino tricycle drivers in the informal economy [26]. They support the need for financial literacy, savings orientation, debt management, and livelihood support programs.

TABLE X. COPING STRATEGIES

Coping Strategy	All f	All %	TODA f	TODA %
Borrow money	73	48.3	20	43.5
Ask relatives for help	46	30.5	25	54.3
Endure without resources	11	7.3	6	13.0
Keep children from school	4	2.6	0	0.0
Others	11	7.3	1	2.2

D. Health, Education, and Service Needs

The third objective was to identify health, education, and service needs. Frequency and ranking were used to determine the common social service needs of respondents.

1) Health needs: Frequency and ranking were used to determine common social-service needs (Table XI). The leading health concern in both groups was cough, colds, or fever (106; 70.2% overall and 30; 65.2% in the transport subgroup). High blood pressure was more pronounced among transport members (12; 26.1%) than in the full sample (21.2%). This is notable because tricycle drivers are routinely exposed to heat, road dust, pollution, fatigue, and long hours, conditions that plausibly elevate cardiovascular and respiratory risk. The results support preventive services such as blood-pressure monitoring, respiratory-health education, dental check-ups, diabetes screening, and wellness activities.

TABLE XI. REPORTED HEALTH CONCERNS

Rank	Health Concern	All f	All %	TODA f	TODA %
1	Cough, colds, or fever	106	70.2	30	65.2
2	Toothache	34	22.5	9	19.6
3	High blood pressure	32	21.2	12	26.1
4	Diabetes	13	8.6	5	10.9
5	Others	7	4.6	3	6.5
6	Skin disease	5	3.3	1	2.2
7	Lung disease	3	2.0	1	2.2

2) Educational needs: Books were the most common educational need in both groups (84; 55.6% overall and 27; 58.7% in the transport subgroup), as shown in Table XII. Educational support thus remained a relevant household concern, and especially so for transport households whose income was generally low and unstable. Assistance may include book donation, school supplies, tutorial assistance, digital-literacy orientation, and scholarship information drives.

TABLE XII. EDUCATIONAL SUPPORT NEEDS

Rank	Educational Need	All f	All %	TODA f	TODA %
1	Books	84	55.6	27	58.7
2	Others	26	17.2	9	19.6
3	Radio	18	11.9	4	8.7
4	Reading materials	5	3.3	0	0.0
4	Educational toys	5	3.3	2	4.3
6	Computer	4	2.6	4	8.7
6	Video camera	4	2.6	1	2.2

3) Public service access: The most frequently identified services overall were elementary education, day-care center, tanod patrol, health center, and cemented road (Table XIII). For the transport subgroup, day-care center, tanod patrol, elementary education, police monitoring, cemented road, and curfew were highly reported. Awareness of existing services did not eliminate household-level needs in health, education, livelihood, and income. The university can therefore act as a linkage institution, connecting members to barangay services, health centers, road-safety programs, livelihood agencies, and social-welfare offices, an approach aligned with calls for meaningful public involvement of underserved transport communities [9], [24], [25].

TABLE XIII. COMMON PUBLIC SERVICES AND FACILITIES IDENTIFIED

Rank	Public Service / Facility	All f	All %	TODA f	TODA %
1	Elementary education	120	79.5	39	84.8
2	Day-care center	115	76.2	44	95.7
3	Tanod patrol	93	61.6	40	87.0
4	Health center	87	57.6	24	52.2
5	Cemented road	77	51.0	38	82.6
6	Basketball court	73	48.3	13	28.3
7	Supplemental feeding	67	44.4	24	52.2
8	Trained community health workers	64	42.4	22	47.8
9	Curfew	63	41.7	38	82.6
10	Police monitoring	61	40.4	39	84.8

E. Ranked Priority Community Needs

Because the dataset used multiple-response and binary-coded items, the priority of each need was measured by the endorsement rate (the proportion of respondents who indicated it), with denominators of 151 for all respondents and 46 for the

transport subgroup (Table XIV). The highest-ranked area for the full sample was transport-oriented support (0.894), driven by the many respondents linked to transport work, tricycle use, road-related indicators, or formal membership; among the transport subgroup this reached 1.000.

TABLE XIV. RANKED PRIORITY COMMUNITY NEEDS

Rank (All)	Priority Area	All f	Endorsement Rate (All)	Rank (TODA)	TODA f	Endorsement Rate (TODA)
1	Transport-oriented support	135	0.894	1	46	1.000
2	Public-service linkage / access	133	0.881	1	46	1.000
3	Health and wellness support	126	0.834	4	39	0.848
4	Educational support	121	0.801	3	40	0.870
5	Livelihood and financial stability	120	0.795	4	39	0.848
6	Skills training and diversification	108	0.715	7	30	0.652
7	Environmental and climate-related support	94	0.623	6	36	0.783

Public-service linkage also ranked highly (0.881 overall; 1.000 for the transport subgroup). Because these items reflected services and facilities identified by respondents, the result should be read as an opportunity for stronger linkage rather than as evidence that all services were absent. Health and wellness ranked third overall and remained important for the transport subgroup, consistent with the high frequency of respiratory illness, hypertension, dental problems, and diabetes.

Educational support ranked fourth overall but higher within the transport subgroup, indicating strong education-related needs—especially for books and learning materials—among lower-income transport households; this mirrors community-survey evidence that educational access is a recurrent social-determinant need [28]. Livelihood and financial stability also emerged as a major priority, supported by low income, borrowing, reliance on relatives, lack of capital, and negative income-expenditure gaps, with the transport subgroup showing stronger vulnerability than the full sample. Skills training and environmental support ranked lower but remained relevant; environmental and climate-related support was especially salient for the transport subgroup, whose members work outdoors and are directly exposed to extreme heat and changing weather.

F. Artificial Intelligence-Assisted Framework and Decision Support

The fifth objective was to translate the cleaned and coded dataset into an Artificial Intelligence-assisted framework. The resulting framework (Fig. 3) shows how technology can support a more responsive extension program. Instead of using survey results only for manual reporting, the workflow transforms the dataset into a decision-support layer that can identify which respondents need livelihood assistance, which households require educational support, which health concerns to prioritize, and which transport-related issues require targeted intervention.

For the transport subgroup, the assisted analysis is especially useful because members share common conditions as transport workers. It distinguishes their needs from the general population by detecting patterns in income, occupation, coping strategies, health risks, road-related concerns, and environmental exposure, allowing the university to design interventions specifically for them. The framework also extends into implementation: it can support a beneficiary database, recommend training modules, identify priority households, monitor participation, and evaluate whether the program improved income, health awareness, skills, or access to services. In this sense, Artificial Intelligence becomes a continuing tool for planning, implementation, monitoring, and evaluation.

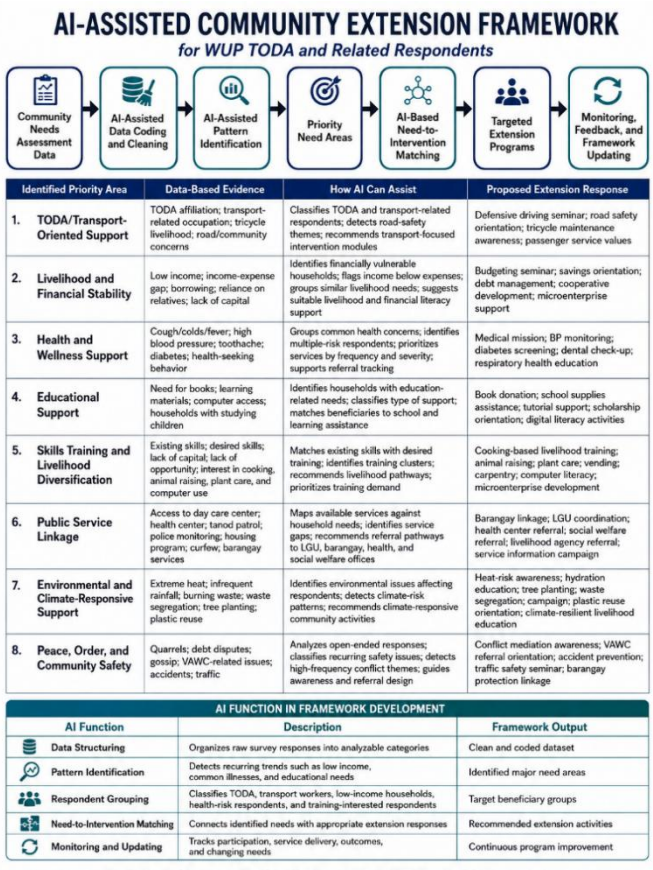


Fig. 3. Artificial Intelligence-assisted community extension framework for the partner community and transport subgroup.

V. CONCLUSION

The Community Needs Assessment dataset of 151 respondents revealed substantial needs in livelihood, health, education, public-service access, transportation support, skills development, and environmental awareness. While all respondents were analyzed, the results emphasized the transport subgroup, which formed a distinct priority sector composed entirely of male tricycle drivers with lower average income and higher livelihood vulnerability.

Profile results showed a largely middle-aged population with household and income responsibilities, and tricycle driving as the dominant livelihood within the transport subgroup. Income and expenditure results indicated financial pressure, particularly for transport members, whose average income was lower than the overall sample. Borrowing and reliance on relatives were common coping strategies, pointing to the need for financial literacy, savings orientation, cooperative development, and livelihood assistance. Health and education were also major concerns, supporting regular medical missions, blood-pressure monitoring, diabetes screening, dental check-ups, respiratory-health education, wellness programs, and educational assistance such as book donations, school supplies, tutorial support, scholarship orientation, and digital-literacy activities.

The ranked priorities indicated that the most important extension areas were transport-oriented support, public-service linkage, health and wellness, educational support, livelihood and financial stability, skills training, and environmental support. Wesleyan University-Philippines should therefore develop an integrated, data-driven extension program for its transport and related partner communities, potentially including defensive-driving seminars, road-safety orientation, tricycle-maintenance awareness, livelihood training, financial literacy, health services, educational assistance, skills development, and barangay or local-government referral support.

Several limitations should be acknowledged. First, the study was confined to a single university partner community, so the findings describe that community and are not statistically generalizable to other settings. Second, the sample was modest (151 respondents, of whom 46 belonged to the transport subgroup), which limits subgroup-level statistical power and the stability of some estimates. Third, the analysis relied on self-reported survey data, which are subject to recall and social-desirability biases, and several income and expenditure fields were incomplete, reducing the number of valid cases for those computations. Fourth, the design was descriptive and cross-sectional; it characterizes needs at one point in time and does not establish causal relationships or change over time. Finally, the Artificial Intelligence component was assistive rather than fully automated, and its clustering and coding outputs depended on researcher review and on the chosen features and parameters.

These limitations point to several directions for future research. Larger and, where feasible, probability-based or multi-community samples would improve representativeness and allow stronger subgroup comparisons. Longitudinal designs could track whether interventions improve income, health awareness, skills, or service access over time. Mixed methods—interviews and focus group discussions—could validate and enrich the quantitative findings. Methodologically, supervised predictive models and more advanced text-embedding or language-model pipelines could be benchmarked against the present descriptive approach, and a full Artificial Intelligence-based decision-support system could be developed and validated for community needs assessment and extension planning.

Finally, the study concludes that Artificial Intelligence can strengthen community extension planning by organizing raw survey data, identifying patterns, grouping respondent populations, ranking priority needs, matching needs with suitable interventions, and supporting monitoring and evaluation. The originality of the work lies in documenting this assistive workflow transparently and showing how it adds value beyond conventional descriptive reporting. It is recommended that the university adopt the proposed framework to make extension programs more targeted, evidence-based, and responsive, and that future researchers validate the findings through interviews, focus group discussions, and the development of an Artificial Intelligence-based decision-support system.

REFERENCES

[1] K. S. Hartmann, N. R. Liburt, and K. Malinowski, “Rutgers equine science center industry needs assessment survey 2016,” Journal of Equine Veterinary Science, vol. 48, pp. 1–8, 2017.

[2] Equine Science Center, “Needs assessment survey,” NFO Plog Research, New Brunswick, NJ, USA, 2002.

[3] K. L. Martinson, T. Bartholomay, K. P. Anderson, C. Skelly, and E. Greene, “Effective evaluation of equine Extension programs,” Journal of Equine Veterinary Science, vol. 32, pp. 616–619, 2012.

[4] K. Martinson, M. Hathaway, J. H. Wilson, B. Gilkerson, P. R. Peterson, and R. Del Vecchio, “University of Minnesota horse owner survey: Building an equine extension program,” Journal of Extension, vol. 44, pp. 1–8, 2006.

[5] A. L. Jaqueth, M. Hathaway, D. N. Catalano, N. C. Linders, R. Mottet, and K. L. Martinson, “Using web-based surveys to explore equine industry practices and future research needs,” Journal of Equine Veterinary Science, vol. 83, art. no. 102822, 2019.

[6] C. Bolwell, D. Gray, and J. Reid, “Identifying the research information needs of the racing and breeding industries in New Zealand: Results of an online survey,” Journal of Equine Veterinary Science, vol. 33, pp. 690–696, 2013.

[7] E. Blumenberg and M. Smart, “Getting by with a little help from my friends… and family: Immigrants and carpooling,” Transportation, vol. 37, pp. 429–446, 2010.

[8] H. Creger, J. Espino, and A. S. Sanchez, Mobility Equity Framework: How to Make Transportation Work for People, 2018.

[9] A. Karner and R. A. Marcantonio, “Achieving transportation equity: Meaningful public involvement to meet the needs of underserved communities,” Public Works Management & Policy, vol. 23, no. 2, pp. 105–126, 2018.

[10] J. D. Landis, “Minority travel disparities and residential segregation: Evidence from the 2017 National Household Travel Survey,” Transportation Research Part D, vol. 112, art. no. 103455, 2022.

[11] K. Manaugh, M. G. Badami, and A. M. El-Geneidy, “Integrating social equity into urban transportation planning,” Transport Policy, vol. 37, pp. 167–176, 2015.

[12] S. R. Arnstein, “A ladder of citizen participation,” Journal of the American Institute of Planners, vol. 35, no. 4, pp. 216–224, 1969.

[13] J. E. Innes, “Consensus building: Clarifications for the critics,” Planning Theory, vol. 3, no. 1, pp. 5–20, 2004.

[14] J. E. Innes and D. E. Booher, “Reframing public participation: Strategies for the 21st century,” Planning Theory & Practice, vol. 5, no. 4, pp. 419–436, 2004.

[15] G. Rowe and L. J. Frewer, "Evaluating public-participation exercises: A research agenda," Science, Technology, & Human Values, vol. 29, no. 4, pp. 512–556, 2004.

[16] C. S. Slotterback and M. Lauria, "Building a foundation for public engagement in planning," Journal of the American Planning Association, vol. 85, no. 3, pp. 183–187, 2019.

[17] T. Webler, S. Tuler, and R. O. B. Krueger, "What is a good public participation process? Five perspectives from the public," Environmental Management, vol. 27, pp. 435–450, 2001.

[18] R. Likert, "A technique for the measurement of attitudes," Archives of Psychology, 1932.

[19] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2021.

[20] A. Agustinez and C. Brown, Disadvantaged Communities Nomenclature Within the State of California: Findings and Conclusions, 2022.

[21] K. L. Einstein and D. M. Glick, "Does race affect access to government services?" American Journal of Political Science, vol. 61, no. 1, pp. 100–116, 2017.

[22] K. L. Einstein, D. Glick, L. Godinez Puig, and M. Palmer, "Still muted: The limited participatory democracy of Zoom public meetings," Urban Affairs Review, vol. 59, no. 4, pp. 1279–1291, 2023.

[23] D. Stempowski, "Counting every voice: Understanding hard-to-count and historically undercounted populations," Census.gov, 2024.

[24] D. Karas, "Highway to inequity: The disparate impact of the interstate highway system on poor and minority communities in American cities," New Visions for Public Affairs, vol. 7, pp. 9–21, 2015.

[25] D. Niemeier, R. Grattet, and T. Beamish, "'Blueprinting' and climate change: Regional governance and civic participation in land use and transportation planning," Environment and Planning C, vol. 33, no. 6, pp. 1600–1617, 2015.

[26] E. A. Onsay, K. C. Baltar, and J. F. Rabajante, "Predicting dual poverty in informal transportation: A machine learning and econometrics approach," 2025, ScienceDirect art. no. S305060772500025X.

[27] E. A. Onsay, J. Alinsunurin, and J. F. Rabajante, "Optimizing machine learning algorithms for multidimensional poverty prediction in the Philippines," SN Business & Economics, vol. 5, no. 10, pp. 1–40, 2025, doi: 10.1007/s43546-025-00922-8.

[28] K. Chansiri, J. S. McCrae, K. Ortega Courtney, and D. Cappello, "A machine learning approach to healthcare needs and barriers using the 100% Community Survey of access to SDOH services," Frontiers in Public Health, vol. 13, art. no. 1659322, 2025, doi: 10.3389/fpubh.2025.1659322.

[29] M. Hossain, L. Jäckering, C. Mullally, and P. Winters, "Poverty prediction and targeting over time and space: Evidence from Nigeria," Applied Economic Perspectives and Policy, vol. 47, no. 3, pp. 1191–1208, 2025, doi: 10.1002/aepp.13515.

[30] P. Verme, "Predicting poverty," The World Bank Economic Review, 2024, doi: 10.1093/wber/lhae044.

[31] J. T. Mjaaland et al., "Scalable and consistent few-shot classification of survey responses using text embeddings," arXiv preprint arXiv:2508.19836, 2025.

A Comparative Analysis of Decision Tree, Random Forest, and Logistic Regression Models in Predicting Business Readiness for Digital Technology Integration

Gloria M. Ducut
College of Technology and Business, Pangasinan State University, Asingan Campus, Pangasinan, Philippines

***Abstract*—This study compared the performance of Decision Tree, Random Forest, and Logistic Regression models in predicting business readiness for digital technology integration using survey data from 400 business respondents in Pangasinan, Philippines. The analysis utilized variables related to technology utilization, perceived helpfulness, willingness to integrate technology, and challenges encountered in adopting digital marketing, e-commerce, and digital payment technologies. Business readiness was operationalized from respondents' willingness scores and classified into Ready and Less Ready categories. Descriptive results revealed very low technology utilization despite high perceived helpfulness, indicating a gap between awareness of digital technologies and their actual adoption. Lack of awareness emerged as the most frequently reported barrier, followed by data security concerns. Using an 80:20 train test split, the machine learning models achieved moderate predictive performance, with Decision Tree and Random Forest attaining the highest accuracy of 63.75%. Random Forest produced the best overall performance, achieving the highest weighted F1 score and demonstrating a more balanced classification capability than the other models. The findings highlight the potential of machine learning as a decision support tool for assessing business readiness and generating evidence-based insights that can support digital transformation planning, technology adoption strategies, and capacity building initiatives among businesses.**

Keywords*—Business readiness; digital technology integration; machine learning

I. Introduction

The rapid advancement of digital technologies has significantly transformed the business landscape, compelling organizations to adopt innovative tools and strategies to remain competitive. Digital transformation has become a critical driver of organizational growth, enabling businesses to improve operational efficiency, strengthen customer engagement, expand market reach, and enhance overall performance [1], [2]. As organizations increasingly integrate digital technologies such as social media platforms, e-commerce systems, and digital payment solutions into their operations, the concept of business readiness has gained considerable attention. Business readiness reflects an organization's preparedness to adopt, utilize, and sustain digital technologies through the availability of technological resources, organizational capabilities, employee competencies, and leadership support [3], [4].

Despite the recognized benefits of digital transformation, many organizations continue to face challenges in technology adoption. Small and medium-sized enterprises (SMEs), in particular, often encounter barriers related to limited awareness, inadequate digital skills, financial constraints, resistance to change, and insufficient organizational support [3],[5]. Consequently, understanding the factors that influence business readiness has become increasingly important for supporting successful digital transformation initiatives. Previous studies have emphasized that organizations with higher levels of digital readiness are better positioned to exploit digital opportunities, improve business processes, and achieve sustainable growth in dynamic business environments [1], [2], [4].

Theoretical perspectives on technology adoption provide valuable insights into the determinants of business readiness. The Technology Acceptance Model (TAM) suggests that perceived usefulness and perceived ease of use influence an individual's intention to adopt technology [6]. Similarly, the Unified Theory of Acceptance and Use of Technology (UTAUT) highlights the importance of performance expectancy, facilitating conditions, and organizational support in shaping technology adoption behavior [7]. Guided by these theoretical perspectives, technology utilization, perceived helpfulness, willingness to integrate technology, and challenges encountered are considered key determinants of organizational readiness for digital technology integration.

Recent developments in machine learning have created new opportunities for analyzing complex datasets and generating predictive insights for decision-making. Machine learning classification techniques have been widely applied in text classification, pattern recognition, predictive analytics, and decision-support systems because of their ability to identify hidden relationships and classify observations based on multiple predictor variables [8]-[11]. Various machine learning approaches, including neural networks, support vector machines, deep learning models, and ensemble learning methods, have demonstrated strong performance in classification tasks across diverse application domains [12]-[15],[18]-[21]. Furthermore, machine learning has been successfully applied to organizational and operational problems such as human resource management, production optimization, and decision-support systems, highlighting its potential for business analytics and strategic planning [16], [17], [22].

Among the commonly used classification algorithms, Decision Tree, Random Forest, and Logistic Regression have attracted considerable attention because they provide a balance between predictive performance and interpretability [23], [24], [25]. Decision Tree models generate transparent decision rules that facilitate understanding of classification outcomes, while Random Forest improves predictive accuracy through ensemble learning and reduced overfitting [25]. Logistic Regression remains a widely used statistical classification method capable of estimating the probability of categorical outcomes and explaining relationships between predictor variables and target classes [23]. These characteristics make the three algorithms suitable for analyzing survey-based datasets involving organizational and behavioral variables.

Although digital readiness and digital transformation have been extensively investigated in information systems and organizational research, the application of machine learning classification techniques for predicting business readiness remains limited. Furthermore, comparative evaluations of Decision Tree, Random Forest, and Logistic Regression models using survey-based organizational data are scarce. This limitation restricts the availability of data-driven tools that can support digital transformation planning and readiness assessment.

Addressing this gap, the present study compares the predictive performance of Decision Tree, Random Forest, and Logistic Regression models in classifying business readiness for digital technology integration. Using indicators related to technology utilization, perceived helpfulness, willingness to integrate technology, and challenges encountered, the study aims to identify the most suitable classification model and generate evidence-based insights that may support digital transformation planning and technology adoption strategies among businesses.

II. Methodology

A. Research Design

This study employed a quantitative predictive research design to evaluate business readiness for digital technology integration using machine learning techniques. Specifically, the predictive performance of Decision Tree, Random Forest, and Logistic Regression models was compared using survey-based data to determine the most accurate and reliable classification algorithm. The study focused on technology adoption in three key areas: digital marketing and social media technologies, e-commerce and online sales platforms, and digital payment solutions. Data were collected through a structured questionnaire that measured respondents' technology utilization, perceived helpfulness of digital tools, willingness to integrate technology into business operations, and challenges encountered during adoption. These variables served as predictors of business readiness and formed the basis for developing and evaluating the machine learning models, thereby providing a data-driven approach to assessing organizational preparedness for digital transformation.

B. Data Source

The dataset used in this study was derived from survey responses collected from business owners and operators in Pangasinan, Philippines. The structured questionnaire gathered information on the adoption and utilization of digital technologies, challenges encountered during technology use, willingness to integrate digital tools into business operations, and perceptions regarding the usefulness of technology in enhancing business performance. Specifically, the survey covered digital marketing technologies, including social media platforms (Facebook, Instagram, TikTok, and YouTube), blogs, Mailchimp, Search Engine Optimization (SEO), Pay-Per-Click (PPC) advertising, Content Management Systems (CMS), marketing analytics tools, and influencer marketing platforms. The dataset also included the use of e-commerce platforms such as Lazada, Shopee, Zalora, and Shein, as well as digital payment solutions including GCash, PayMaya, DragonPay, online banking transfers, and Coins.ph. These responses served as the primary data source for developing and evaluating the machine learning models used to predict business readiness for digital technology integration.

C. Respondents of the Study

The respondents of the study consisted of business owners, managers, and authorized business representatives from various enterprises in Pangasinan, Philippines. These individuals were selected because they possessed sufficient knowledge of their organizations' digital technology practices, including the adoption and utilization of digital marketing tools, e-commerce platforms, and digital payment systems. As key decision-makers or individuals directly involved in business operations, they were considered reliable sources of information regarding technology usage, challenges encountered during digital adoption, and the willingness of their businesses to integrate additional digital technologies. Only complete and valid responses were included in the analysis. Prior to model development, the dataset underwent a data cleaning process to identify and address incomplete, inconsistent, or erroneous entries, ensuring the quality and reliability of the data used for machine learning classification.

D. Variables of the Study

The dependent variable in this study was business readiness for digital technology integration. Since the survey instrument did not directly measure readiness, a readiness index was operationalized using respondents' willingness to integrate technology. This approach is grounded in the Technology Acceptance Model (TAM), which identifies behavioral intention as a key determinant of technology adoption, and the Unified Theory of Acceptance and Use of Technology (UTAUT), which emphasizes the influence of performance expectancy and facilitating conditions on technology acceptance and use [6], [7]. Accordingly, willingness to integrate technology was used as a proxy indicator of readiness because it reflects an organization's intention and preparedness to adopt digital technologies when enabling conditions are present. Guided by these theoretical frameworks, technology utilization, perceived helpfulness, and challenges encountered were selected as predictor variables because they represent important dimensions of technology acceptance, perceived value, and adoption barriers that may influence readiness for digital technology integration [6], [7]. These variables served as the primary inputs for the machine learning models used to classify

business readiness and examine the factors associated with digital technology adoption.

TABLE I. VARIABLES USED IN THE MACHINE LEARNING MODELS

Variable	Description	Type	Coding
Technology Utilization	Level of use of digital technologies in business operations	Ordinal	1 = Not Utilized, 2 = Low Utilization, 3=Moderate Utilization, 4 = High Utilization, 5 = Heavily Utilized
Perceived Helpfulness	Perceived usefulness of digital technologies for business operations	Ordinal	1 = Not Helpful, 2 = Slightly Helpful, 3=Moderately Helpful, 4 = Helpful, 5 = Extremely Helpful
Willingness to Integrate Technology	Intention to adopt and integrate digital technologies	Ordinal	1 = Not Willing, 2 = Slightly Willing, 3=Moderately Willing, 4 = Very Willing, 5 = Extremely Willing
Challenges Encountered	Presence of barriers to technology adoption	Binary	0 = No, 1 = Yes
Business Readiness	Target variable derived from willingness scores	Categorical	Less Ready (< 3.00), Ready (≥ 3.00)

Table I outlines the variables used in the machine learning models for predicting business readiness for digital technology integration. Technology utilization, perceived helpfulness, and challenges encountered served as the predictor variables, while business readiness was used as the target variable. Since the survey instrument did not contain a direct measure of readiness, business readiness was derived from respondents' willingness to integrate technology and classified into two categories: Less Ready and Ready. The selection of these variables was guided by the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT), which emphasize the roles of perceived usefulness, behavioral intention, and facilitating conditions in influencing technology adoption.

E. Data Preparation

Prior to model development, the survey dataset underwent data preparation and preprocessing to ensure its quality, consistency, and suitability for machine learning analysis. The dataset was examined for missing values, duplicate records, inconsistent responses, and invalid entries, with incomplete responses excluded and minor missing values addressed through appropriate imputation techniques. Because the survey data primarily consisted of categorical and ordinal responses, these variables were transformed into numerical values to facilitate model training. Ordinal responses related to technology utilization, willingness to integrate technology, and perceived helpfulness were encoded using a five-point scale, while challenge-related variables were converted into binary values, with 1 indicating the presence of a challenge and 0 indicating its absence. Following data transformation, the dataset was reviewed to remove variables with excessive missing values, limited variability, or minimal analytical relevance. Business readiness was subsequently derived from respondents' willingness to integrate technology and classified into Ready and Less Ready categories based on the established readiness threshold. To avoid target leakage, willingness variables were used solely for the construction of the target variable and were not included as predictors during model training. Consequently, technology utilization, perceived helpfulness, and challenges encountered were retained as the final predictor variables used in the machine learning models.

F. Data Splitting

Following data preprocessing, the dataset was partitioned into training and testing sets to support model development and evaluation. The training dataset was used to train the Decision Tree, Random Forest, and Logistic Regression models, while the testing dataset was reserved for assessing their predictive performance using previously unseen observations. An 80:20 train test split was employed, with 80% of the data allocated for model training and 20% reserved for testing. This proportion is widely used in supervised machine learning studies because it provides sufficient data for model learning while maintaining an independent dataset for performance evaluation. Prior to model training, the distribution of the target variable was examined to identify potential class imbalance issues. Where necessary, appropriate measures were considered to ensure that the classification models were not overly influenced by the majority class, thereby improving the reliability and robustness of the prediction results.

G. Machine Learning Models

This study employed three supervised machine learning classification algorithms, namely Decision Tree, Random Forest, and Logistic Regression, to predict business readiness for digital technology integration. The Decision Tree model was selected because of its ability to generate clear and interpretable decision rules, making it useful for understanding how different predictor variables contribute to readiness classification. Random Forest was included as an ensemble learning technique that combines multiple decision trees to improve predictive performance and reduce the risk of overfitting. Logistic Regression was utilized as a statistical classification approach that estimates the probability of membership in a particular readiness category and provides interpretable relationships between predictor variables and classification outcomes. The inclusion of these three models allowed for a comprehensive comparison of classification performance, enabling the identification of the most suitable algorithm for predicting business readiness using survey-based data.

H. Model Training and Testing

Each classification model was trained using the training dataset, with technology utilization, perceived helpfulness, and challenges encountered serving as the predictor variables. The target variable was the business readiness classification derived from respondents' willingness to integrate technology. Following model training, the Decision Tree, Random Forest, and Logistic Regression models were evaluated using the testing dataset, which contained observations not included during the training process. This approach enabled the assessment of each model's predictive performance on previously unseen data and provided an indication of its ability to generalize beyond the

training dataset. By separating model training from model evaluation, the study minimized the risk of overestimating model performance and ensured a more objective comparison of the classification algorithms.

I. Evaluation Metrics

The predictive performance of the Decision Tree, Random Forest, and Logistic Regression models was evaluated using standard classification metrics, including accuracy, precision, recall, F1 score, and the confusion matrix. Accuracy measured the overall proportion of correctly classified observations, while precision assessed the proportion of correct predictions within a given readiness category. Recall measured the ability of the model to correctly identify actual readiness classifications, and the F1 score provided a balanced assessment of precision and recall. The confusion matrix offered a detailed summary of correct and incorrect classifications across the readiness categories, enabling a more comprehensive evaluation of model behavior. Collectively, these metrics were used to compare the performance of the three machine learning algorithms and identify the most effective model for predicting business readiness for digital technology integration.

J. Comparative Analysis

Following model evaluation, a comparative analysis was conducted to assess the performance of the Decision Tree, Random Forest, and Logistic Regression models using accuracy, precision, recall, F1 score, and confusion matrix results. The model demonstrating the strongest overall performance across these metrics was identified as the most suitable approach for predicting business readiness for digital technology integration. In addition to predictive performance, model interpretability was also considered to support practical decision making. Accordingly, the models were compared not only in terms of classification accuracy but also in their ability to provide insights into how technology utilization, perceived helpfulness, and challenges encountered influence business readiness. This approach enabled the evaluation of both the predictive effectiveness and practical applicability of the machine learning models.

K. Ethical Statement

Participation was voluntary, and informed consent was obtained from all respondents prior to data collection. Respondents were assured that their responses would remain confidential and would be used solely for academic and research purposes. No personally identifiable information was collected or reported. All data were analyzed and presented in aggregated form to protect respondent privacy and comply with applicable ethical and data protection guidelines.

III. Results and Discussion

A. Descriptive Analysis of Business Readiness and Technology Adoption

The analysis was based on the raw survey dataset consisting of 400 valid respondent records collected from businesses in Pangasinan, Philippines. The survey measured key dimensions of digital technology adoption, including technology utilization, challenges encountered, willingness to integrate technology, and perceived helpfulness of digital marketing, e-commerce, and digital payment tools. These variables were aligned with the study's conceptual framework and served as the primary inputs for the machine learning models. Since the dataset did not contain a direct measure of business readiness, a readiness index was operationalized using respondents' willingness to integrate digital technologies. Consistent with technology adoption theories that identify behavioral intention as a strong predictor of actual technology use [6], [7], readiness was derived from the average willingness scores across the technology categories included in the survey. Respondents with an average willingness score of 3.00 or higher were classified as Ready, while those with scores below 3.00 were classified as Less Ready. This classification approach provided a practical and theoretically grounded method for generating the target variable used in the machine learning analysis.

TABLE II. Technology Utilization, Willingness to Integrate, and Perceived Helpfulness by Technology Category.

Technology Category	Mean Utilization	Mean Willingness	Mean Helpfulness
Digital Marketing and Social Media	1.12	2.79	4.38
E-commerce and Online Sales	1	3.03	4.56
Digital Payment Solutions	1.56	3.25	4.31

Table II presents the technology utilization, willingness to integrate, and perceived helpfulness of digital technologies across three major categories: digital marketing and social media, e-commerce and online sales, and digital payment solutions. The results revealed that digital payment solutions recorded the highest mean utilization score (1.56), indicating that these technologies were more widely adopted than the other categories. This finding suggests that businesses tend to prioritize digital technologies that provide immediate operational benefits, such as faster transactions, improved payment convenience, and enhanced customer service. The relatively high utilization of GCash and other digital payment platforms reflects the growing acceptance of cashless transactions and the practical value of digital payment systems in supporting day-to-day business activities.

In contrast, e-commerce and online sales technologies exhibited the lowest mean utilization score (1.00), indicating limited adoption among the surveyed businesses. Despite this low level of utilization, e-commerce received the highest perceived helpfulness score (4.56), demonstrating that respondents recognized its potential to expand market reach, improve customer access, and increase business opportunities. This disparity between perceived helpfulness and actual utilization suggests that businesses may be aware of the benefits of e-commerce but face barriers that prevent its full adoption. Similar observations have been reported in digital transformation studies, which emphasize that awareness of technology benefits does not automatically translate into implementation due to constraints related to resources, digital competencies, organizational readiness, and facilitating conditions [1]-[5].

Digital payment solutions also achieved the highest willingness to integrate score (3.25), followed by e-commerce

technologies (3.03). These findings indicate that businesses are more prepared to adopt technologies that offer visible and immediate advantages. According to the Technology Acceptance Model, perceived usefulness significantly influences behavioral intention toward technology adoption [6]. Likewise, the Unified Theory of Acceptance and Use of Technology highlights the role of performance expectancy and facilitating conditions in shaping adoption behavior [7]. The relatively higher willingness scores observed for digital payment and e-commerce technologies suggest that businesses perceive these tools as beneficial and potentially valuable for improving operational efficiency and competitiveness.

The findings further reveal a notable gap between technology utilization and perceived helpfulness across all categories. Although respondents generally viewed digital technologies as useful, actual adoption remained limited. This pattern suggests that readiness for digital transformation is influenced by factors beyond perceived usefulness alone. Previous studies have identified organizational capabilities, digital literacy, leadership support, technological infrastructure, and strategic readiness as important determinants of successful digital transformation [1]–[5]. Consequently, initiatives aimed at improving awareness, digital skills, and access to technology resources may help bridge the gap between positive perceptions and actual technology adoption. Overall, the results indicate that while businesses recognize the value of digital technologies, further support and capacity-building efforts are needed to enhance readiness and promote broader digital technology integration.

TABLE III. MOST UTILIZED TECHNOLOGIES

Rank	Technology	Mean Utilization
1	GCash	3.71
2	Facebook	2.44
3	Reels	1.14
4	Online Banking Transfer	1.07
5	TikTok	1.04

Table III shows the most utilized technologies. The most utilized tool was GCash, with a mean utilization score of 3.71, followed by Facebook with 2.44. This suggests that respondents were most familiar with technologies that directly support payment transactions and customer communication. Most other tools had very low utilization, with several platforms recording a mean score of 1.00, including Lazada, Shopee, Zalora, Shein, PayMaya, DragonPay, and Coins.ph.

TABLE IV. HIGHEST WILLINGNESS TO INTEGRATE TECHNOLOGIES

Rank	Technology	Mean Willingness
1	GCash	4.64
2	Facebook	4.15
3	Online Banking Transfer	3.8
4	YouTube	3.73
5	Shopee / Lazada	3.5

Table IV presents the technologies with the highest willingness to integrate among the respondents. The respondents showed the highest willingness to integrate GCash, with a mean willingness score of 4.64, followed by Facebook with 4.15. This means that digital payment and social media tools were the most acceptable technologies among the respondents. On the other hand, lower willingness scores were observed for Coins.ph, Content Management Systems, Marketing Analytics, and Pay-Per-Click Advertising, suggesting that these tools may require more awareness, training, or support before businesses become ready to adopt them.

TABLE V. CHALLENGES ENCOUNTERED

Challenge	Frequency	Percentage
Lack of awareness	6,040	65.65%
Data security concerns	1,514	16.46%
Inadequate skills, training, and support	446	4.85%
Resistance to change among employees	443	4.82%
Lack of financial resources	439	4.77%
Connectivity or infrastructure issues	293	3.18%
Complexity of the technology/system	25	0.27%

[a] **Note:** Frequencies and percentages were calculated from the total number of coded challenge responses because respondents were allowed to select multiple challenges.

Table V illustrates the challenges encountered by businesses in adopting digital technologies. Lack of awareness emerged as the most frequently reported barrier, accounting for 65.65% of all coded responses, indicating that many businesses may have limited knowledge of available digital tools and their potential benefits for business operations. Data security concerns ranked second at 16.46%, suggesting that issues related to privacy, trust, and the protection of digital transactions also influence technology adoption decisions. In contrast, the low frequency of responses related to technology complexity indicates that businesses do not necessarily perceive digital tools as difficult to use. Rather, the findings suggest that limited awareness and concerns regarding security represent the primary obstacles to digital technology integration. These results highlight the importance of digital literacy initiatives, awareness campaigns, and cybersecurity education programs to improve business readiness and encourage wider adoption of digital technologies.

TABLE VI. BUSINESS READINESS CLASSIFICATION

Readiness Classification	Frequency	Percentage
Less Ready	240	60.00%
Ready	160	40.00%
Total	400	100.00%

Table VI presents the business readiness classification of the respondents. Based on the established readiness threshold, 240 respondents (60.00%) were classified as Less Ready, while 160 respondents (40.00%) were classified as Ready for digital technology integration. These findings indicate that although respondents generally recognized the benefits and usefulness of digital technologies, the majority had readiness scores below the threshold required for classification as Ready. The results suggest that positive perceptions of digital technologies do not necessarily translate into readiness for adoption, highlighting the need for greater awareness, capacity building, and support

mechanisms to facilitate digital technology integration among businesses.

TABLE VII. MACHINE LEARNING MODEL RESULTS

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
Decision Tree	0.6375	0.625	0.6375	0.6129
Random Forest	0.6375	0.6325	0.6375	0.6341
Logistic Regression	0.6125	0.6356	0.6125	0.6163

Table VII provides the performance of the machine learning models using an 80:20 train-test split. Decision Tree and Random Forest achieved the highest accuracy of 63.75%, while Random Forest obtained the highest weighted F1 score (0.6341), indicating the best overall classification performance. Logistic Regression achieved a slightly lower accuracy of 61.25% but recorded the highest weighted precision (0.6356). The moderate predictive performance of the models reflects the complexity of business readiness, which is influenced by multiple organizational and contextual factors beyond those captured in the survey. Despite relying on perception-based variables, the results demonstrate that machine learning, particularly Random Forest, can effectively identify readiness patterns and support data-driven decision-making for digital technology integration. Future studies may improve predictive performance by incorporating additional variables related to organizational capabilities, digital literacy, internet accessibility, and prior technology experience [1]-[5], [25].

TABLE VIII. CONFUSION MATRIX RESULTS

Model	Actual Less Ready Correctly Predicted	Actual Less Ready Predicted Ready	Actual Ready Predicted Less Ready	Actual Ready Correctly Predicted
Decision Tree	40	8	21	11
Random Forest	35	13	16	16
Logistic Regression	28	20	11	21

Table VIII summarizes the confusion matrix results of the three machine learning models. The Decision Tree model correctly classified most Less Ready businesses but showed limited effectiveness in identifying Ready businesses, indicating a tendency toward the majority class. Logistic Regression demonstrated better performance in identifying Ready businesses; however, it produced a higher number of false positive classifications, which reduced its overall accuracy. In contrast, the Random Forest model achieved a more balanced classification performance by accurately identifying both Less Ready and Ready businesses. Although its accuracy was comparable to that of the Decision Tree model, its higher weighted F1 score reflected a better balance between precision and recall across the two readiness categories. These findings suggest that Random Forest provided the most reliable and practical predictive performance for assessing business readiness for digital technology integration, making it the most suitable model for supporting data driven decision making and digital transformation planning.

B. Discussion

The findings revealed a notable gap between the perceived usefulness of digital technologies and their actual utilization among businesses in Pangasinan. Although respondents generally recognized the value of digital technologies in improving business operations, customer engagement, and market reach, actual technology utilization remained relatively low. This disparity suggests that positive perceptions alone do not guarantee technology adoption. According to the Technology Acceptance Model (TAM), perceived usefulness influences behavioral intention, which subsequently affects technology use [6]. Similarly, the Unified Theory of Acceptance and Use of Technology (UTAUT) highlights the importance of facilitating conditions and organizational support in translating intention into actual adoption behavior [7]. Thus, while businesses may acknowledge the benefits of digital technologies, existing barriers may prevent them from fully integrating these tools into their operations.

Among the technology categories examined, digital payment solutions demonstrated the highest level of adoption and readiness, largely driven by the widespread use of GCash and other mobile payment platforms. This finding suggests that businesses are more inclined to adopt technologies that offer immediate operational benefits, including transaction convenience, improved efficiency, and reduced cash handling risks. Previous studies have shown that organizations prioritize digital technologies that provide clear business value and contribute directly to operational performance and competitiveness [1],[2]. In contrast, the relatively low utilization of e-commerce platforms indicates that many businesses have not fully embraced online selling despite recognizing its usefulness. This finding may reflect gaps in digital competencies, awareness, organizational capabilities, and readiness, which have been identified as important determinants of successful digital transformation [3]-[5].

The challenge analysis further supports this interpretation. The predominance of lack of awareness as a reported barrier suggests that informational and knowledge-related constraints may be more significant than purely financial or technical limitations. This finding underscores the importance of digital literacy programs, technology demonstrations, business mentoring, and capacity building initiatives to enhance readiness for digital technology integration. Previous research has emphasized that digital readiness extends beyond technological infrastructure and includes leadership support, organizational capabilities, employee competencies, and the ability to adapt to changing business environments [3]-[5]. Consequently, efforts aimed at improving awareness and digital skills may substantially increase the likelihood of successful technology adoption among businesses.

With respect to predictive modeling, Random Forest emerged as the best-performing classification algorithm, achieving the highest overall accuracy and weighted F1 score among the evaluated models. The superior performance of Random Forest may be attributed to its ensemble learning mechanism, which combines multiple decision trees to improve predictive accuracy and reduce overfitting [25]. Although Decision Tree and Logistic Regression provided greater interpretability and clearer explanations of classification

outcomes, Random Forest demonstrated a stronger ability to accurately classify both Ready and Less Ready businesses. These findings suggest that machine learning techniques can effectively predict business readiness using variables related to technology utilization, perceived helpfulness, and challenges encountered. The ability of the models to identify readiness patterns demonstrates the usefulness of machine learning as a decision support tool for assessing digital transformation readiness and technology adoption behavior. Consistent with previous studies, machine learning algorithms have proven effective in classification, prediction, and decision support applications across various organizational and operational contexts [16], [17], [22]. Although the predictive performance observed in this study was moderate, the results indicate that technology adoption characteristics can provide meaningful insights into business readiness and may support evidence-based planning, capacity building, and digital transformation initiatives among businesses.

The results further indicate that business readiness is influenced by factors beyond those captured in the present dataset. Although willingness to integrate technology served as a practical proxy for readiness, digital transformation literature recognizes readiness as a multidimensional construct that includes organizational resources, technological infrastructure, leadership commitment, digital competencies, and strategic orientation [1]-[5]. While TAM and UTAUT support the use of behavioral intention as a predictor of technology adoption [6], [7], future research should develop and validate more comprehensive readiness frameworks that incorporate technological, organizational, financial, and human resource dimensions. Such an approach would strengthen the conceptual measurement of readiness while improving the predictive capability of machine learning models.

Several limitations should be considered when interpreting the findings. First, the study employed a single 80:20 train-test split for model evaluation, which may influence performance estimates because classification results can vary depending on how the data are partitioned. Future studies should employ more robust validation approaches, such as k-fold cross-validation, stratified k-fold validation, or repeated random subsampling, to generate more stable and reliable performance estimates. Second, although the dataset consisted of 400 valid responses, only a subset was used for testing, which may affect the generalizability of the reported performance metrics. Third, the study focused on businesses within a specific geographical context, potentially limiting the applicability of the findings to other regions or business sectors. Future studies should therefore utilize larger and more diverse datasets to improve model robustness and external validity.

Another limitation is that feature importance analysis was not performed for the Random Forest model. Although Random Forest achieved the highest predictive performance, the relative contribution of individual predictor variables was not quantified. Consequently, the study focused primarily on model comparison and predictive accuracy rather than model explainability. Future research should incorporate feature importance analysis and model interpretation techniques such as permutation importance and SHAP (Shapley Additive Explanations) to identify the variables that most strongly influence business readiness. Such analyses would provide deeper insights into technology adoption behavior and support the development of more targeted and evidence based digital transformation strategies.

Guo (2025) [26] compared Decision Tree, Logistic Regression, AdaBoost, and Random Forest models in predicting corporate credit default and reported that the improved Random Forest model achieved superior predictive performance, with an accuracy of 94.2%, recall of 93.5%, and F1-score of 92.8%. The study attributed the strong performance of Random Forest to its ensemble structure, which enhances classification accuracy while reducing the risk of overfitting commonly associated with single-tree models. Although the predictive performance observed in the present study was more moderate due to the complexity and perception-based nature of business readiness data, the superior performance of Random Forest similarly demonstrates its capability to capture complex relationships among predictor variables and generate more balanced classifications than Decision Tree and Logistic Regression models. These findings further support the suitability of Random Forest as a practical decision-support tool for organizational assessment, technology adoption studies, and business analytics applications.

The findings of the present study are consistent with recent machine learning research that highlights the effectiveness of ensemble learning methods for classification tasks. Guo (2025) compared Decision Tree, Logistic Regression, AdaBoost, and Random Forest models in predicting corporate credit default and reported that the improved Random Forest model achieved superior predictive performance, with an accuracy of 94.2%, recall of 93.5%, and F1-score of 92.8%. The study attributed the strong performance of Random Forest to its ensemble structure, which enhances classification accuracy while reducing the risk of overfitting commonly associated with single-tree models. Although the predictive performance observed in the present study was more moderate due to the complexity and perception-based nature of business readiness data, the superior performance of Random Forest similarly demonstrates its capability to capture complex relationships among predictor variables and generate more balanced classifications than Decision Tree and Logistic Regression models. These findings further support the suitability of Random Forest as a practical decision-support tool for organizational assessment, technology adoption studies, and business analytics applications.

IV. Conclusion and Recommendations

A. Conclusion

The study concluded that business readiness for digital technology integration among businesses in Pangasinan remains in the developmental stage. Although respondents generally perceived digital technologies as highly beneficial for improving business operations, customer engagement, and market expansion, actual utilization levels remained relatively low, indicating a gap between awareness of benefits and practical adoption. Digital payment solutions, particularly GCash, exhibited the highest levels of utilization and willingness to integrate, suggesting that businesses are more likely to adopt technologies that provide immediate and tangible operational advantages. In contrast, e-commerce platforms and several digital marketing technologies remained underutilized despite

their perceived usefulness. Lack of awareness emerged as the primary barrier to adoption, followed by concerns regarding data security, highlighting the need for digital literacy initiatives, technology training, and business support programs. Among the machine learning models evaluated, Random Forest demonstrated the strongest predictive performance and was identified as the most suitable model for classifying business readiness, while Decision Tree and Logistic Regression provided valuable interpretability and baseline comparisons. Further, the findings demonstrate that machine learning can serve as an effective tool for predicting business readiness and generating evidence-based insights that support digital transformation planning and technology adoption strategies among businesses.

B. Recommendations

Businesses should strengthen digital awareness and skills through training and digital literacy programs, particularly in digital marketing, e-commerce, and digital payment systems. Given that lack of awareness was the primary barrier to adoption, local government units, business support organizations, and technology providers should collaborate in providing technical assistance and capacity-building initiatives. Future studies should utilize larger datasets, incorporate additional organizational and contextual variables, and explore advanced machine learning algorithms to improve the prediction of business readiness for digital technology integration. Future studies should also employ k-fold cross-validation and feature importance analysis to improve model robustness and enhance the interpretability of machine learning predictions.

ACKNOWLEDGMENT

The author gratefully acknowledges the Pangasinan State University for its support in the conduct and completion of this study. The author also utilized artificial intelligence-assisted tools to improve language clarity, readability, and editorial quality during manuscript preparation. These tools were not used to generate research data, perform data analysis, interpret results, or draw conclusions. All analysis, interpretations, and final manuscript content were reviewed, verified, and approved by the author, who assumes full responsibility for the accuracy and integrity of the work.

REFERENCES

[1] F. Li, "The digital transformation of business models in the creative industries: A holistic framework and emerging trends," Technovation, vol. 92–93, pp. 102012, 2020, doi: 10.1016/j.technovation.2017.12.004.

[2] M. Verhoef, T. Broekhuizen, Y. Bart, A. Bhattacharya, J. Dong, N. Fabian, and M. Haenlein, "Digital transformation: A multidisciplinary reflection and research agenda," Journal of Business Research, vol. 122, pp. 889–901, 2021, doi: 10.1016/j.jbusres.2019.09.022.

[3] A. A. AlNuaimi, K. Singh, S. Ren, V. Budhwar, and D. Vorobyev, "Mastering digital transformation: The nexus between leadership, agility, and digital readiness," Journal of Business Research, vol. 145, pp. 636–648, 2022, doi: 10.1016/j.jbusres.2022.03.038.

[4] M. North, T. Aramburu, and C. Lorenzo, "Promoting digitally enabled growth in SMEs: The role of digital readiness and organizational capabilities," Technology in Society, vol. 72, pp. 102178, 2023, doi: 10.1016/j.techsoc.2022.102178.

[5] A. Troise, E. Corvello, A. Ghobadian, and O. O'Regan, "How can SMEs successfully navigate VUCA environment: The role of agility in the digital transformation era," Technological Forecasting and Social Change, vol. 174, pp. 121227, 2022, doi: 10.1016/j.techfore.2021.121227.

[6] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

[7] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quarterly, vol. 27, no. 3, pp. 425–478, 2003, doi: 10.2307/30036540.

[8] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Berlin/Heidelberg, Germany: Springer, 2012, pp. 163–222.

[9] C. C. Aggarwal and C. X. Zhai, Mining Text Data. Berlin/Heidelberg, Germany: Springer, 2012.

[10] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese news text classification based on machine learning algorithm," in Proc. Int. Conf. Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 2, 2018, pp. 48–51.

[11] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," WSEAS Transactions on Computers, vol. 4, no. 8, pp. 966–974, 2005.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[13] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural Processing Letters, vol. 9, pp. 293–300, 1999.

[14] M. Heidarysafa, K. Kowsari, D. E. Brown, K. Jafari Meimandi, and L. E. Barnes, "An improvement of data classification using random multimodel deep learning (RMDL)," International Journal of Machine Learning and Computing, vol. 8, pp. 298–310, 2018.

[15] K. Kowsari, M. Heidarysafa, D. E. Brown, K. Jafari Meimandi, and L. E. Barnes, "RMDL: Random multimodel deep learning for classification," in Proc. Int. Conf. Information System and Data Mining, Lakeland, FL, USA, Apr. 9–11, 2018, doi: 10.1145/3206098.3206111.

[16] N. Pourkhodabakhsh, M. M. Mamoudan, and A. Bozorgi-Amiri, "Effective machine learning, meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover," Applied Intelligence, 2022, doi: 10.1007/s10489-022-04294-6.

[17] Y. Li, S. Carabelli, E. Fadda, D. Manerba, R. Tadei, and O. Terzo, "Machine learning and optimization for production rescheduling in Industry 4.0," International Journal of Advanced Manufacturing Technology, vol. 110, no. 9–10, pp. 2445–2463, 2020, doi: 10.1007/s00170-020-05850-5.

[18] O. A. Montesinos-López, A. Montesinos-López, and J. Crossa, "Fundamentals of artificial neural networks and deep learning," in Multivariate Statistical Machine Learning Methods for Genomic Prediction, Cham, Switzerland: Springer International Publishing, 2022, pp. 379–425, doi: 10.1007/978-3-030-89010-0_10.

[19] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26–38, 2017, doi: 10.1109/MSP.2017.2743240.

[20] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in Proc. Twenty-Ninth AAAI Conf. Artificial Intelligence, Austin, TX, USA, Jan. 25–30, 2015, vol. 333, pp. 2267–2273.

[21] S. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," IEEE Transactions on Neural Networks, vol. 3, no. 5, pp. 683–697, 1992, doi: 10.1109/72.159058.

[22] H. Lee, N. Aydin, Y. Choi, S. Lekhavat, and Z. Irani, "A decision support system for vessel speed decision in maritime logistics using weather archive big data," Computers & Operations Research, vol. 98, pp. 330–342, 2018, doi: 10.1016/j.cor.2017.06.005.

[23] S. Domínguez-Almendros, N. Benítez-Parejo, and A. Gonzalez-Ramirez, "Logistic regression models," Allergologia et Immunopathologia, vol. 39, no. 5, pp. 295–305, 2011, doi: 10.1016/j.aller.2011.05.002.

[24] J.-S. Chou and J. Thedja, "Metaheuristic optimization within machine learning-based classification system for early warnings related to geotechnical problems," Automation in Construction, vol. 68, pp. 65–80, 2016, doi: 10.1016/j.autcon.2016.03.015.

[25] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[26] K. Guo, "Corporate credit default prediction based on machine learning," in Proceedings of the 2025 International Conference on Economic Management and Big Data Application (ICEMBDA 2025), 2025.

Canthus-Scaled 468-Landmark FaceMesh Framework for Pupillary Distance Estimation Using Nested AutoML Calibration

Mohd Izzuddin Mohd Tamrin[1], Sherzod Turaev[2], Takumi Sase[3],
Mohd Zulfaezal Che Azemin[4*], Tengku Mohd Tengku Sembok[5]
Kulliyyah of ICT, International Islamic University Malaysia, Gombak, Kuala Lumpur, Malaysia[1, 3, 5]
Department of Computer Science and Software Engineering-College of Information Technology,
United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates[2]
Integrated Omics Research Group-Kulliyyah of Allied Health Sciences,
International Islamic University Malaysia, Kuantan, Malaysia[4]

Abstract*—Pupillary distance (PD) is an important ocular measurement for optical dispensing and vision-related applications, but standard MediaPipe FaceMesh outputs do not provide true pupil-centre or iris-boundary landmarks when only the 468-landmark representation is available. This study proposes a canthus-scaled 468-landmark framework for estimating PD using facial landmarks and Malay young-adult normative palpebral fissure width. A dataset of 44 subjects was used, where each record contained ground-truth PD and 1,404 coordinate values representing 468 FaceMesh landmarks in three dimensions. Since true pupil centres were unavailable, medial and lateral canthus landmarks were used to construct eye-centre proxies and to compute a subject-specific millimetre scale. A direct canthus-scaled proxy was first evaluated as a deterministic baseline, after which canthus-scaled geometric features were used in a nested AutoML calibration framework. Model development used repeated nested cross-validation, with an outer repeated 5-fold design and an inner 4-fold model-selection loop. The direct proxy achieved a mean absolute error (MAE) of 4.26 mm and showed systematic overestimation. The calibrated nested AutoML model improved performance, achieving a subject-level MAE of 3.510 mm, root mean squared error of 4.22 mm, a bias of −0.08 mm, and 75.0% of predictions within ±5 mm. The calibrated nested AutoML model improved overall error and reduced systematic bias compared with the direct canthus-scaled proxy. However, the Bland–Altman limits of agreement remained wide, indicating that the proposed method should be interpreted as an approximate proxy-based estimation approach rather than a substitute for clinical pupillometer- or ruler-based PD measurement. The framework is most relevant for research settings or datasets where only standard 468-landmark FaceMesh data are available, and iris-refined landmarks are absent.

Keywords*—Pupillary distance; MediaPipe FaceMesh; facial landmarks; canthus scaling; palpebral fissure width; AutoML; nested cross-validation; computer vision

I. Introduction

Pupillary distance (PD), commonly discussed as interpupillary distance (IPD), is the horizontal distance between the centres of the two pupils and is an important measurement in optometry, ophthalmic dispensing, binocular vision assessment, and the fitting of ophthalmic lenses. Inaccurate PD measurement may affect optical centration and can contribute to visual discomfort, especially when lenses are prescribed or manufactured using incorrect centration parameters. Conventional PD measurement is commonly performed using clinical instruments such as pupillometers, autorefractors, or manual PD rulers. However, these approaches may require trained personnel, dedicated equipment, or direct clinical access, which motivates the development of low-cost and automated alternatives [1], [2].

Recent studies have investigated smartphone- and computer-vision-based methods for estimating PD or related ocular measurements. Jung and Chu compared modern IPD measurement methods, including clinical and mobile application approaches, showing the growing interest in non-contact digital PD estimation [1]. Han et al. evaluated smartphone applications for IPD measurement, further supporting the feasibility of mobile-based measurement but also highlighting the need for accuracy validation [2]. Zhang et al. proposed a computer-vision method for IPD and pupil-height measurement using ensemble regression trees and segmentation-based techniques, demonstrating that automated ocular measurement remains an active area of applied vision research [3]. Similarly, pupil-centre detection has been investigated using convolutional neural networks and webcam-based systems, indicating that accurate pupil localization is technically challenging but central to reliable ocular measurement [4].

MediaPipe FaceMesh has become a widely used framework for real-time facial landmark extraction because it can estimate dense three-dimensional facial landmarks from ordinary RGB images. The standard FaceMesh model estimates 468 facial landmarks, making it attractive for lightweight and non-invasive facial analysis applications [5]. However, the standard 468-landmark representation does not include true iris-boundary or pupil-centre landmarks. MediaPipe Iris extends the FaceMesh framework by estimating additional iris and eye-contour landmarks, enabling more direct iris- and pupil-related measurements from a single RGB camera [6], [7]. In datasets where only the 468 FaceMesh

*Corresponding author.

landmarks are available, PD cannot be measured directly from true pupil centres; instead, proxy-based estimation using periocular landmarks is required.

Population-specific anthropometric information is important when facial or ocular landmarks are converted from image-space distances into metric units. Ngeow and Aljunid established craniofacial anthropometric norms for young adult Malaysian Malays aged 18–25 years, providing population-relevant reference measurements for craniofacial and orbital regions [8]. Othman et al. further established three-dimensional facial soft-tissue morphology for adult Malay subjects using stereophotogrammetry, reinforcing the importance of population-specific facial measurements in Malaysian samples [9]. Additional Malaysian and regional studies on palpebral fissure and canthal measurements show that periocular dimensions vary across sex and ethnic groups, supporting the use of appropriate normative references when converting facial landmark distances into estimated millimetres [10]–[13].

Although iris-based scaling is preferable when iris landmarks are available, many existing landmark datasets contain only 468 FaceMesh landmarks. In such cases, canthus-based scaling provides a practical alternative. The medial and lateral canthus landmarks can be used to estimate palpebral fissure width, and a normative palpebral fissure width value can then be used as a scale reference. This allows the distance between canthus-derived eye-centre proxies to be converted into an estimated PD value. However, because the eye-centre proxies are not anatomical pupil centres, a direct normative conversion may introduce systematic bias. Machine-learning calibration can therefore be used to learn the relationship between canthus-scaled facial landmark features and ground-truth PD.

The key limitation addressed in this study is that many existing facial-landmark datasets contain only the standard 468 MediaPipe FaceMesh landmarks and do not include iris-boundary or anatomical pupil-centre landmarks. Under this constraint, PD cannot be measured directly from the true pupil centres. This study, therefore, investigates a two-stage framework: first, a geometric canthus-scaled proxy is constructed using medial and lateral canthus landmarks and Malay young-adult normative palpebral fissure width; second, a predictive calibration model is trained to reduce systematic proxy error using nested cross-validation. The contributions of this study are: 1) a 468-landmark canthus-scaled proxy method for approximate PD estimation, 2) an evaluation of direct geometric estimation versus calibrated prediction, 3) a nested AutoML validation strategy for a small biomedical dataset, and 4) an explicit discussion of the clinical and deployment limitations of the proxy-based approach.

II. Related Work

Previous work on pupillary distance estimation can be grouped into four main areas: conventional clinical measurement, smartphone- and computer-vision-based ocular measurement, facial landmark and iris-based estimation, and validation methods for small-sample prediction models. Conventional PD measurement is commonly performed using pupillometers, autorefractors, or manual rulers, which remain more appropriate for clinical dispensing because they are designed to measure ocular centration directly. However, these approaches require equipment, trained personnel, or clinical access, motivating low-cost computer-vision alternatives.

Smartphone and computer-vision methods have shown promise for non-contact PD and related ocular measurements, but their accuracy depends strongly on pupil or iris localization, camera geometry, subject pose, and calibration strategy. Methods that use direct pupil or iris detection are conceptually closer to anatomical PD measurement, whereas approaches based only on facial landmarks must rely on surrogate geometric features. MediaPipe FaceMesh provides dense facial landmarks and is attractive for lightweight facial analysis, but the standard 468-landmark model does not include true pupil-centre or iris-boundary landmarks. MediaPipe Iris extends this representation with additional iris-related landmarks and is therefore preferable when iris landmarks are available.

In contrast, the present study focuses on the more restricted setting where only 468 FaceMesh landmarks are available. Under this condition, canthus landmarks provide a practical periocular reference because they are visible in the standard model and can be linked to palpebral fissure width. Anthropometric scaling is also important because image-space distances must be converted into millimetres. Population-specific normative values may reduce scaling error compared with arbitrary scaling, but they cannot account fully for inter-subject periocular variation. Therefore, this study treats canthus scaling as an approximate population-level calibration rather than a subject-specific physical calibration.

Finally, because the dataset is small and model selection is part of the modelling process, nested cross-validation is used to reduce optimistic performance estimation. This is important because using cross-validation both for model selection and final error estimation can introduce biased performance estimates [14]. In addition, the reporting of the proposed prediction framework was informed by TRIPOD+AI guidance for regression and machine learning-based prediction models, particularly in relation to transparent description of the data, predictors, validation procedure, performance metrics, and limitations [15].

III. Methodology

The proposed framework consists of two conceptually distinct stages. Stage 1 is geometric proxy estimation, in which canthus landmarks are used to estimate palpebral fissure width, construct eye-centre proxies, and convert landmark distances into estimated millimetres. Stage 2 is predictive calibration, in which canthus-scaled geometric features are used to learn the relationship between proxy geometry and reference PD. The first stage is deterministic and interpretable, whereas the second stage is a data-driven correction layer. The calibrated model should therefore be interpreted as a correction of a canthus-derived proxy, not as a direct measurement of anatomical pupil-centre distance.

A. Dataset and Landmark Representation

This study used a subject-level facial landmark dataset for pupillary distance (PD) estimation. Each record contained a subject identifier, a ground-truth PD value in millimetres, and

1,404 coordinate values representing 468 MediaPipe FaceMesh landmarks in three-dimensional form. The landmark values were arranged as 468 x-coordinates, 468 y-coordinates, and 468 z-coordinates. The dataset contained 44 usable subjects after retaining records with valid subject identifiers, ground-truth PD values, and complete landmark coordinates.

The standard 468-landmark FaceMesh representation does not include anatomical pupil centres or iris-boundary landmarks. Therefore, this study estimated PD using canthus-derived eye-centre proxies and a canthus-based normative scaling procedure. The complete methodological workflow is illustrated in Fig. 1.

B. Landmark Pre-processing

The uploaded landmark matrix was first parsed into a structured coordinate tensor. For each subject, the 1,404 landmark values were reshaped into a 468×3 matrix, where each landmark was represented by x, y, and z coordinates. Records of missing or non-numeric landmark values were excluded to ensure complete geometric calculations.

The following FaceMesh landmarks were used as the primary ocular reference points:

L_33and L_133: right-eye lateral and medial canthus region landmarks.

L_263and L_362: left-eye lateral and medial canthus region landmarks.

These landmarks were selected because they are available in the 468-landmark FaceMesh model and correspond to the canthal region required for palpebral fissure width estimation.

C. Canthus-Based Normative Scaling

Since no physical calibration object or iris landmarks were available, image-space landmark distances were converted into millimetres using a population-level canthus-based normative scale. A normative palpebral fissure width (PFW) of 29.100 mm was used as the scaling constant for the Malay young-adult population.

For each subject, the right and left image-space palpebral fissure widths were calculated as:

$$PFW_R = d(L_{33}, L_{133}) \quad (1)$$

$$PFW_L = d(L_{263}, L_{362}) \quad (2)$$

The subject-specific image-space palpebral fissure width was then computed as:

$$PFW_{image} = \frac{PFW_R + PFW_L}{2} \quad (3)$$

The millimetre-per-landmark-unit scale factor was calculated as:

$$S = \frac{PFW_{normative}}{PFW_{image}} \quad (4)$$

where, PFW_normative=29.100mm.

This scale factor was applied to convert selected landmark distances from landmark units into estimated millimetres. This scaling approach assumes that the selected normative palpebral fissure width is representative of the target population. However, it does not account for subject-specific anatomical variation in eyelid aperture, canthal position, sex-related differences, or facial pose. To make this limitation explicit, the subject-specific scale factor and right-left palpebral fissure asymmetry were retained as scale quality-control variables.

D. Eye-Centre Proxy Estimation

Because true pupil centres were not available in the 468-landmark representation, this study did not attempt to measure anatomical PD directly. Instead, each eye was represented by a canthus-derived geometric proxy defined as the midpoint between the medial and lateral canthus landmarks. The right and left eye-centre proxies were calculated as:

$$E_R = \frac{L_{33} + L_{133}}{2} \quad (5)$$

$$E_L = \frac{L_{263} + L_{362}}{2} \quad (6)$$

The direct canthus-scaled proxy PD was then calculated as:

$$PD_{proxy} = d(E_R, E_L) \times S \quad (7)$$

This direct proxy was used as a deterministic baseline for comparison with the machine-learning-based prediction model. This midpoint is expected to differ from the true anatomical pupil centre because the pupil position depends on gaze, eyelid aperture, ocular anatomy, and camera perspective. Therefore, the direct canthus-scaled estimate was treated only as a deterministic baseline and not as a clinical PD measurement.

E. Feature Engineering

Canthus-scaled geometric features were constructed from the 468-landmark coordinate representation. The engineered feature groups included:

- Eye-corner midpoint distance, computed from the distance between the right and left canthus-derived eye-centre proxies.
- Eye-contour centroid distance, computed from the centroid of available eye-contour landmarks.
- Anatomical pairwise landmark distances, including ocular, facial-width, mouth-width, and vertical face-distance proxies.
- Scale quality-control features, including the subject-specific scale factor and the percentage asymmetry between right and left palpebral fissure widths.
- Full landmark representations, including all 468 landmarks and subject-level normalized landmark features.

All distance-based geometric features were converted into estimated millimetres using the subject-specific canthus scale factor. Because the number of subjects was limited, the high-dimensional all-landmark representation was treated cautiously. The candidate model set, therefore, included low-dimensional eye-region and anatomical-distance feature sets, scale quality-control variables, and regularized models. The all-landmark feature set was included only as one candidate representation within the nested selection procedure, not as the default model.

F. Direct Proxy Baseline

A direct canthus-scaled PD proxy was calculated for each subject using only the eye-centre proxy distance and the normative canthus scale factor. This method did not require model training and served as an interpretable baseline for evaluating whether a simple normative conversion was sufficient for PD estimation.

To evaluate whether the added complexity of AutoML calibration was justified, simple statistical calibration baselines were added. These included a linear correction model using the direct canthus-scaled proxy as the predictor and a polynomial correction model using first- and second-order terms of the direct proxy. These models were evaluated using the same outer cross-validation structure as the AutoML framework, so that calibration performance could be compared under the same validation conditions.

G. Nested AutoML Prediction Model

A nested cross-validation framework was used to train and evaluate machine-learning models while avoiding data leakage. The outer loop used repeated 5-fold cross-validation with 10 repeats. In each outer fold, subjects in the test split were held out and were not used during model selection or hyperparameter tuning. The inner loop used 4-fold cross-validation on the training split to select the optimal feature set, estimator, and hyperparameters.

The AutoML candidate set included baseline and regression-based models using canthus-scaled feature sets. Candidate models included dummy regression, linear eye-region estimators, Ridge regression, ElasticNet regression, support vector regression, and partial least squares regression. Model selection was based on minimizing mean absolute error in millimetres within the inner cross-validation loop.

For each outer fold, the best-performing model from the inner loop was refitted on the corresponding outer training set and used to predict PD for the held-out subjects. Repeated outer-fold predictions were averaged to obtain one subject-level prediction per participant.

A landmark perturbation analysis was conducted to evaluate sensitivity to plausible localization error. Key canthus landmarks and eye-contour landmarks were perturbed with zero-mean random noise at predefined magnitudes, and the direct proxy and calibrated predictions were recomputed. Performance was summarized using MAE, RMSE, bias, and Bland–Altman limits of agreement. Because the available dataset contained landmark coordinates rather than raw images, image-level effects such as illumination variation, motion blur, occlusion, and head-pose changes could not be fully evaluated. These image-level robustness factors are therefore identified as priorities for future validation.

H. Evaluation Protocol

The direct proxy baseline and nested AutoML predictions were evaluated against ground-truth PD values. The evaluation metrics included mean absolute error, root mean squared error, bias, standard deviation of error, median absolute error, coefficient of determination, Bland–Altman limits of agreement, and the percentage of predictions within ±2 mm and ±5 mm.

Bootstrap resampling was used to estimate confidence intervals for subject-level performance metrics. The final analysis compared the deterministic canthus-scaled proxy with the calibrated nested AutoML model to assess whether learned calibration improved PD estimation from 468-landmark canthus-scaled features.

I. Workflow Summary

The proposed methodology consists of landmark parsing, canthus-based scale estimation, eye-centre proxy calculation, direct proxy PD estimation, canthus-scaled feature construction, nested AutoML calibration, and final evaluation. This process is summarized in Fig. 1.

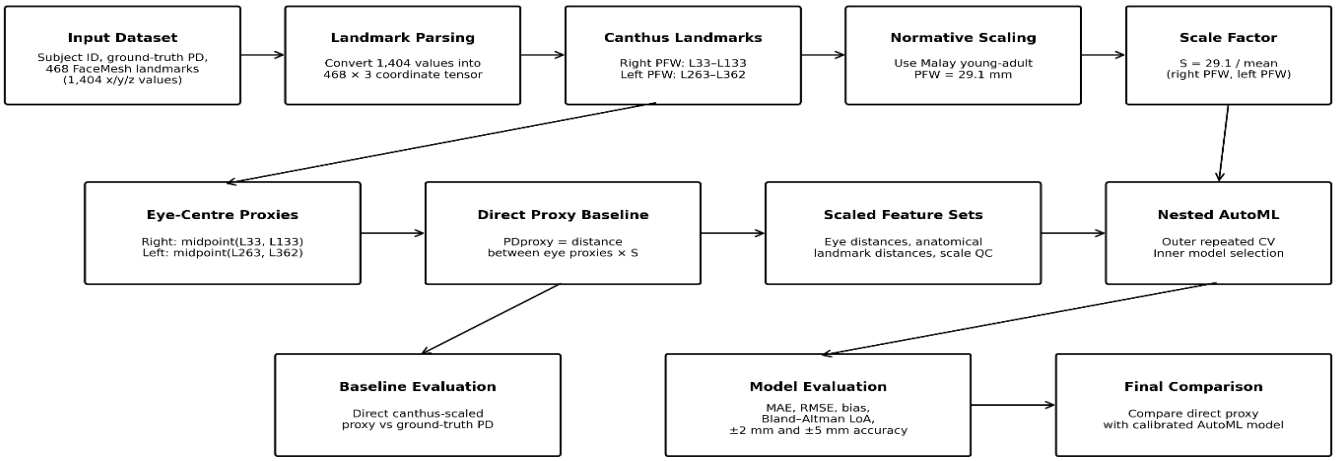


Fig. 1. Overall methodology of the 468-landmark canthus-scaled PD estimation framework.

IV. Results and Discussion

A. Dataset and Experimental Configuration and Acronyms

The final analysis included 44 subjects with complete 468-landmark FaceMesh data and corresponding ground-truth pupillary distance (PD) values. Each subject contributed 1,404 landmark coordinate values, representing 468 landmarks in x, y, and z dimensions. The proposed workflow used a canthus-based normative scaling procedure, where landmarks 33–133 and 263–362 were used to estimate right and left palpebral fissure widths, respectively. The overall experimental configuration is summarized in Table I.

The use of nested cross-validation was appropriate because the dataset was small and model selection was performed during training. In small-sample prediction modelling, non-nested validation may produce unstable or optimistic estimates, whereas nested validation separates model selection from performance estimation [16], [17].

The canthus-based scale factor showed measurable inter-subject variability. The mean scale factor was 1109.74 mm per landmark unit with a standard deviation of 141.41 mm per landmark unit. The mean right-left palpebral fissure width asymmetry was 4.95%, while the maximum asymmetry reached 44.09%. These values indicate that although normative scaling provides a practical conversion from landmark units to millimetres, individual periocular morphology and landmark asymmetry may introduce scaling error.

B. Direct Canthus-Scaled Proxy Baseline

The direct canthus-scaled proxy served as a deterministic baseline. This method estimated PD by calculating the distance between the right and left canthus-derived eye-centre proxies and multiplying it by the subject-specific canthus scale factor. As shown in Table II, the direct proxy achieved an MAE of 4.26 mm and RMSE of 4.91 mm. The bias was +2.83 mm, indicating that the direct proxy generally overestimated the ground-truth PD.

This finding suggests that canthus-based scaling alone is insufficient for accurate PD estimation when true pupil centres are unavailable. The direct proxy is interpretable and simple, but it depends on the assumption that the midpoint between the medial and lateral canthus approximates the pupil centre. Since the 468-landmark representation does not include actual pupil-centre landmarks, systematic geometric bias is expected.

TABLE I. Dataset and Experimental Configuration

Item	Value
N usable subjects	44
Target PD mean (mm)	63.59
Target PD SD (mm)	4.26
Target PD median (mm)	62.50
Target PD min (mm)	55.00
Target PD max (mm)	72.00
Landmark values per subject	1404.00
Interpreted landmarks	468.00
Metric scaling approach	468-landmark canthus-based normative scaling
Normative palpebral fissure width (mm)	29.10
Scale landmarks	Right 33-133 and left 263-362; scale = normative PFW / mean observed PFW
Direct proxy formula	distance(midpoint(33,133), midpoint(263,362)) x canthus scale
Outer CV	Repeated 5-fold CV x 10 repeats
Inner CV	4-fold CV
Primary model selection metric	Mean absolute error (MAE), mm

TABLE II. Performance Comparison Between Direct Proxy and Nested AutoML

Analysis level	MAE (mm)	RMSE (mm)	Bias (mm)	Median AE (mm)	Within +/-2 mm (%)	Within +/-5 mm (%)	LoA lower (mm)	LoA upper (mm)	95% CI
Direct canthus-scaled proxy	4.26	4.91	2.83	4.31	20.45	59.09	-5.12	10.80	
Nested AutoML subject-level	3.51	4.23	-0.08	3.57	31.82	75.00	-8.45	8.30	MAE 2.82-4.21; RMSE 3.47-4.91; Bias -1.34-1.12
Nested AutoML outer-fold	3.61	4.35	-0.08	3.46	32.27	72.05	-8.61	8.46	

TABLE III. AutoML-Selected Candidate Frequency Across Outer Folds

Selected candidate	Selected feature set	Outer folds	Mean inner MAE (mm)	SD inner MAE (mm)	Selection (%)
Eye-centroid estimator	eye_centroid_scaled_mm	23	3.45	0.21	46.00
Eye-corner estimator	eye_corner_scaled_mm	6	3.21	0.35	12.00
Geometry SVR-RBF	geometry_26_scaled_mm	6	3.28	0.17	12.00
Geometry Ridge	geometry_26_scaled_mm	6	3.47	0.22	12.00
Geometry ElasticNet	geometry_26_scaled_mm	5	3.22	0.13	10.00
All-landmarks PLS	all_landmarks_scaled_mm	3	3.39	0.14	6.00
Dummy median baseline	geometry_26_scaled_mm	1	3.396		2.00

C. Nested AutoML Prediction Performance

The nested AutoML model improved the overall prediction accuracy compared with the direct proxy baseline. As reported in Table II, the subject-level nested AutoML prediction achieved an MAE of 3.51 mm, RMSE of 4.22 mm, and bias of −0.08 mm. The proportion of predictions within ±5 mm increased to 75.00%, compared with 59.09% for the direct proxy baseline. The bootstrap 95% confidence interval for subject-level MAE was 2.82–4.21 mm, indicating the expected uncertainty range of the model's average absolute error.

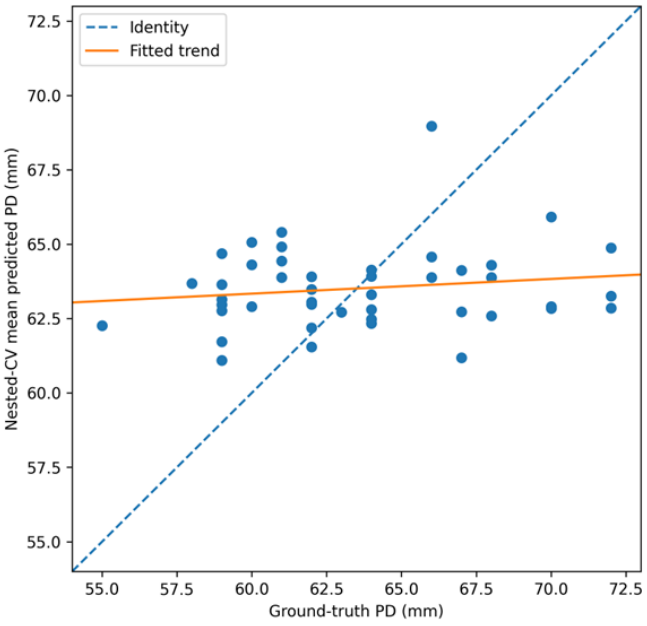


Fig. 2. Predicted versus ground-truth PD using nested AutoML.

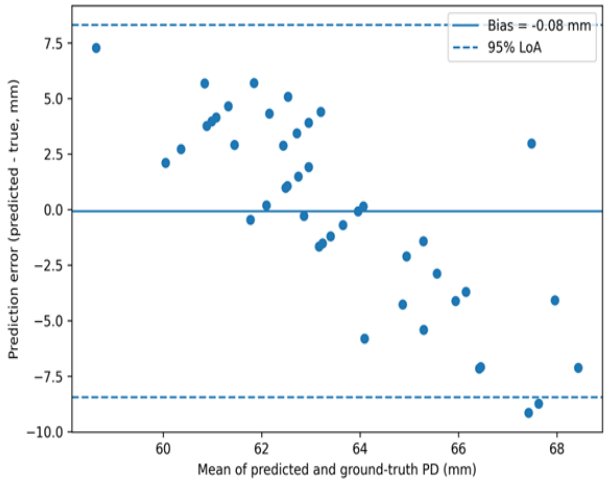


Fig. 3. Bland-Altman analysis of nested AutoML PD estimation.

The predicted-versus-ground-truth plot in Fig. 2 shows that the AutoML predictions were concentrated within a narrower prediction range than the ground-truth PD values. This indicates that the model learned a calibrated central tendency but had limited ability to fully represent the extremes of the PD distribution. This behaviour is common in small-sample regression models, where predictions may shrink toward the mean when the training set contains limited variation [18].

D. Agreement Analysis

The Bland–Altman plot in Fig. 3 provides a clinically interpretable view of agreement between the nested AutoML predictions and the ground-truth PD values. The average bias was close to zero, which indicates that the learned calibration model corrected the systematic overestimation observed in the direct proxy baseline. However, the limits of agreement remained relatively wide, showing that individual-level prediction error may still be meaningful.

Bland–Altman analysis is useful in this context because correlation or fitted trend alone does not fully describe agreement between a new measurement method and a reference method [19]. Therefore, Fig. 3 is more informative than relying only on the scatter plot in Fig. 2, especially for evaluating whether the method is suitable for measurement-related applications.

E. AutoML Model-Selection Behaviour

The AutoML selection results are summarized in Table III. The selection frequencies indicate that simpler eye-region and geometry-based models were selected more often than the full all-landmark model. The eye-centroid estimator was selected in 46.0% of outer folds, whereas the all-landmarks PLS model was selected in only 6.0% of outer folds. This suggests that the nested selection procedure did not consistently favor the highest-dimensional representation. Nevertheless, the small sample size means that feature selection may still be unstable, and the selected model frequencies should be interpreted as exploratory rather than definitive evidence of feature importance.

This selection pattern suggests that interpretable eye-region features were more stable than the full high-dimensional landmark representation. The relatively low selection frequency of the all-landmark PLS model indicates that using all 468 landmarks did not consistently improve model selection. This is methodologically important because the sample size was small relative to the number of potential landmark predictors. Prediction modelling literature recommends caution when the number of predictors is large compared with the number of observations, since overfitting and unstable selection may occur [17], [18].

F. Comparison Between Direct Proxy and Learned Calibration

The comparison between the direct proxy and nested AutoML model shows that the main benefit of AutoML was not simply reducing error magnitude but reducing systematic bias. The direct canthus-scaled proxy was easy to compute but overestimated PD. In contrast, the nested AutoML model used the same canthus-scaled landmark information while learning a calibration relationship between proxy geometry and reference PD. This explains why the AutoML model achieved a much smaller bias while also improving MAE and RMSE.

For the manuscript, the direct proxy should therefore be presented as a baseline rather than the final proposed estimator. The final model should be described as a calibrated 468-

landmark canthus-scaled PD estimation framework. This framing is important because the 468-landmark FaceMesh model does not provide anatomical pupil centres; consequently, the model estimates PD from canthus-derived ocular proxies rather than measuring pupil-centre distance directly.

G. Practical Implications and Limitations

The proposed approach is useful when only standard 468 FaceMesh landmarks are available, and iris landmarks cannot be extracted. It provides a practical method for converting landmark geometry into estimated millimetres using Malay young-adult normative palpebral fissure width and then calibrating the estimate using nested AutoML.

However, several limitations should be considered. First, the method relies on normative palpebral fissure width rather than subject-specific physical calibration. Second, the eye-centre proxy is derived from canthus landmarks and is not equivalent to a true pupil centre. Third, the dataset contained only 44 subjects, so the reported performance should be interpreted as preliminary. Finally, the model should be externally validated on a larger independent Malay undergraduate cohort before being used as a general-purpose PD estimation method.

Overall, the results support the feasibility of canthus-scaled 468-landmark PD estimation, but they also show that direct geometric conversion is not sufficient on its own. Learned calibration through nested AutoML improved bias and overall error while preserving an interpretable landmark-based measurement framework.

H. Clinical Suitability and Agreement Limits

Although the nested AutoML model reduced the average bias to −0.08 mm, the Bland–Altman limits of agreement remained wide, ranging from −8.45 to +8.30 mm. The total agreement width was therefore approximately 16.75 mm. This level of variability is not acceptable for clinical spectacle dispensing, where centration errors of only a few millimetres can be practically important, particularly for progressive or high-powered lenses. Therefore, the proposed framework should not be interpreted as a replacement for clinical PD measurement using a pupillometer, autorefractor, or trained manual measurement. Its current value is as an approximate proxy-estimation method for research use, preliminary screening, or retrospective datasets where only 468 FaceMesh landmarks are available.

I. Interpretation of the Canthus-Derived Eye-Centre Proxy

The central assumption of the geometric stage is that the midpoint between the medial and lateral canthus provides a stable periocular reference. This assumption is useful computationally because the landmarks are available in the standard 468-landmark FaceMesh representation. However, the canthus midpoint is not equivalent to the anatomical pupil centre. Differences between the proxy and the true pupil centre may arise from gaze direction, eyelid shape, palpebral fissure morphology, subject pose, and landmark localization error. The systematic overestimation observed in the direct proxy baseline is consistent with this limitation. The calibrated AutoML model partially reduced this bias, but it cannot fully recover true pupil-centre geometry when the relevant anatomical landmarks are absent.

J. Effect of Periocular Anatomical Variability

The use of Malay young-adult normative palpebral fissure width provides a population-relevant scale reference, but it does not eliminate inter-subject anatomical variability. Subjects with wider or narrower palpebral fissures than the normative value, asymmetric canthal landmarks, or atypical eyelid morphology may receive inaccurate scale factors. The observed scale-factor variability and palpebral fissure asymmetry support this concern. Future work should investigate sex-specific scaling, subject-specific calibration objects, iris-based scaling when available, and exclusion or quality-control criteria for cases with high canthal asymmetry.

K. Small Sample Size and Model Generalizability

The dataset contained 44 subjects, which is small relative to the number of possible landmark-derived predictors. Although repeated nested cross-validation was used to reduce optimism from model selection, it cannot fully remove the risk of unstable feature selection or limited external generalizability. The results should therefore be interpreted as preliminary internal-validation findings. External validation on a larger independent cohort is required before the method can be recommended for broader deployment.

L. Robustness to Real-World Image Variation

The present dataset contained processed landmark coordinates rather than raw images. Therefore, the study could not directly evaluate the effect of illumination, image blur, camera resolution, head pose, gaze direction, eyelid occlusion, or partial landmark failure. These factors may substantially affect practical performance because the proposed framework depends on accurate detection of canthus and eye-contour landmarks. The perturbation analysis provides an initial landmark-level sensitivity assessment, but image-level robustness must be evaluated in future studies using controlled acquisition protocols and raw image data.

V. Conclusion

This study developed and evaluated a canthus-scaled 468-landmark framework for estimating pupillary distance from MediaPipe FaceMesh landmarks. Since the standard 468-landmark representation does not provide true pupil centres or iris-boundary landmarks, the proposed approach used medial and lateral canthus landmarks to construct eye-centre proxies and to derive a subject-specific metric scale based on normative Malay young-adult palpebral fissure width. This enabled pupillary distance estimation from landmark geometry without requiring iris-refined landmarks or a physical calibration object.

The direct canthus-scaled proxy provided a simple and interpretable baseline; however, it was limited by systematic overestimation because canthus-derived midpoints are not equivalent to anatomical pupil centres. The nested AutoML calibration model improved the estimation by learning the relationship between canthus-scaled landmark features and ground-truth pupillary distance. The use of repeated nested cross-validation reduced optimism from model selection and

provided a more reliable estimate of predictive performance for the small dataset.

Overall, the findings indicate that standard 468 FaceMesh landmarks can support approximate proxy-based PD estimation when combined with population-specific canthus scaling and calibration. However, the approach should not be regarded as a direct clinical measurement of pupil-centre distance. The wide limits of agreement, small sample size, reliance on normative scaling, and absence of true iris or pupil landmarks limit immediate clinical applicability. Future work should validate the framework on a larger independent cohort, compare simple calibration models with AutoML models, evaluate robustness to landmark and image-level errors, incorporate sex-specific or subject-specific scaling, and benchmark the 468-landmark approach against iris-refined landmark models and standard clinical PD measurement instruments.

Declaration on Generative AI

ChatGPT 5.5 was used to assist with language editing and to improve clarity during manuscript preparation. The authors take full responsibility for the content, confirm that all scientific interpretations and conclusions are their own, and approve the final version of the manuscript.

Acknowledgment

The authors would like to express their sincere appreciation to Ms. Nor Alya Batrisyia Nor Izani for her valuable assistance during the data collection process. Her support contributed to the successful preparation of the dataset used in this study.

References

[1] Y. R. Jung and B. S. Chu, “A comparative analysis of interpupillary distance measurement techniques evaluation in modern times: From rulers to apps,” Clinical Optometry, vol. 16, pp. 309–316, 2024.

[2] K. D. Han, M. Jaafar, I. M. Stoakes, P. C. Hoopes, and M. Moshirfar, “Comparing the effectiveness of smartphone applications in the measurement of interpupillary distance,” Cureus, vol. 15, no. 7, Art. no. e42744, 2023.

[3] Z. Zhang, H. Xiang, D. Li, and C. Leng, “Measurement method of interpupillary distance and pupil height based on ensemble of regression trees and the BlendMask algorithm,” Applied Sciences, vol. 13, no. 15, Art. no. 8628, 2023.

[4] A. Larumbe-Bergera, G. Garde, S. Porta, R. Cabeza, and A. Villanueva, “Accurate pupil center detection in off-the-shelf eye tracking systems using convolutional neural networks,” Sensors, vol. 21, no. 20, Art. no. 6847, 2021.

[5] A. Ablavatski, I. Grishchenko, Y. Kartynnik, K. Raveendran, and M. Grundmann, “Attention Mesh: High-fidelity face mesh prediction in real-time,” arXiv preprint arXiv:2006.10962, 2020.

[6] Google MediaPipe, “MediaPipe Face Mesh,” Google, 2024.

[7] Google Research, “MediaPipe Iris: Real-time iris tracking & depth estimation,” Google Research Blog, 2020.

[8] W. C. Ngeow and S. T. Aljunid, “Craniofacial anthropometric norms of Malays,” Singapore Medical Journal, vol. 50, no. 5, pp. 525–528, 2009.

[9] S. A. Othman, L. P. Majawit, W. N. W. Hassan, M. C. Wey, and R. M. Razi, “Anthropometric study of three-dimensional facial morphology in Malay adults,” PLOS ONE, vol. 11, no. 10, Art. no. e0164180, 2016.

[10] T. Y. Lu, K. Kadir, W. C. Ngeow, and S. A. Othman, “The prevalence of double eyelid and the 3D measurement of orbital soft tissue in Malays and Chinese,” Scientific Reports, vol. 7, Art. no. 14829, 2017.

[11] V. Packiriswamy, P. Kumar, and K. G. M. Rao, “Photogrammetric analysis of palpebral fissure dimensions and its position in Malaysian South Indian ethnic adults by gender,” North American Journal of Medical Sciences, vol. 4, no. 10, pp. 458–462, 2012.

[12] W. C. Ngeow and S. T. Aljunid, “Craniofacial anthropometric norms of Malaysian Indians,” Indian Journal of Dental Research, vol. 20, no. 3, pp. 313–319, 2009.

[13] N. P. Murray, M. Hunfalvay, and T. Bolte, “The reliability, validity, and normative data of interpupillary distance and pupil diameter using eye-tracking technology,” Translational Vision Science & Technology, vol. 6, no. 4, Art. no. 2, 2017.

[14] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” BMC Bioinformatics, vol. 7, Art. no. 91, 2006.

[15] G. S. Collins, K. G. M. Moons, P. Dhiman, and et al., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, vol. 385, Art. no. e078378, 2024.

[16] G. Varoquaux, “Cross-validation failure: Small sample sizes lead to large error bars,” NeuroImage, vol. 180, pp. 68–77, 2018.

[17] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. Cham, Switzerland: Springer, 2019.

[18] F. E. Harrell Jr., Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. Cham, Switzerland: Springer, 2015.

[19] J. M. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” The Lancet, vol. 327, no. 8476, pp. 307–310, 1986.

A Hybrid CNN-Ensemble Framework for Robust DeepFake Image Detection

Mohammad Alsulami
College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia

***Abstract*—The fast development of deepfake technologies has caused growing concerns related to the authentication of digital media, the integrity of personal identification, and the spreading of disinformation. Therefore, there is a growing need for effective automatic detectors of deepfakes. Meanwhile, existing techniques of deepfake detection encounter a large number of difficulties. First, it is difficult to distinguish the subtle manipulation details of faces. Another problem is poor generalization when applying models to novel datasets or highly realistic, generated synthetic faces. Another issue is low accuracy in the case of imbalanced data, where samples of one class dominate others. Thus, to address those problems, this study proposes a novel hybrid approach based on a combination of deep learning (DL) and conventional machine learning (ML) for detecting deepfake images. More precisely, two pre-trained CNNs (MobileNetV2 and ResNet50) were applied to generate features and classify them via the Random Forest (RF) algorithm. The experiments have been conducted on a benchmark of 6,557 facial images marked as either real or fake. The findings reveal that the MobileNetV2+RF achieved the highest accuracy 99%, followed by MobileNetV2 with 98% and ResNet50 with 97% accuracy. This suggests that the hybrid architecture helps to increase the effectiveness of the solution. A statistical significance test reveals that none of the models' performances differ significantly from each other ($p > 0.05$). Overall, the proposed system demonstrates excellent metrics concerning accuracy and precision.**

***Keywords*—Deep learning models; machine learning classifiers; image processing; data augmentation**

I. Introduction

The rapid evolution in the fields of artificial intelligence (AI), deep learning, and multimedia technologies has transformed the way people interact and engage in creating digital information [1]. Currently, smartphones, laptop computers, and sophisticated computer systems provide individuals with the capacity to develop, disseminate, and access digital multimedia in a faster way by using online platforms. This technological development has enabled many benefits in fields such as entertainment, the educational sector, health care services, and even social media communications. However, they have posed many problems associated with digital security and integrity [2]. Deepfake technology is one of those serious problems.

Deepfakes refer to media that have been modified using digitally manipulated images, videos, or audio that were produced through artificial intelligence techniques [3]. The artificial intelligence used includes generative adversarial networks and autoencoders. Through this technology, highly accurate fake faces and voices have been produced since the technology is capable of mimicking human facial expressions and their voices. Deepfake technology has experienced great progress ever since its invention in 2017 due to the emergence of free and available artificial intelligence technology, the availability of big data online, and the availability of computational power [4].

Misuse of deepfake technology has caused several significant threats to individuals, organizations, and even society as a whole. The uses of deepfakes have been identified to mislead people, shape their opinions politically, engage in identity fraud, conduct financial crimes, and create inappropriate content. Social media platforms like Facebook, Instagram, Youtube have made deepfake videos and other forms of content go viral on the internet in an instant due to the speed at which information can travel over them. On the other hand, deepfake technology is a concern in many fields, including forensics, surveillance, and biometrics [5], [6], [7].

Facial recognition and identification systems based on conventional methods depend heavily on using biometric data associated with the face. In most cases, such algorithms do not cope with the identification of manipulated images of people's faces. They are less efficient with images that have undergone various types of manipulations, such as blur processing, compression, sizing, filtering, and enhancement. Meanwhile, development in the field of image manipulation and AI-based face generation has increased the realism of synthetic faces [8], [9].

To address these issues, recent works in this field have been concentrating on designing techniques for automatically detecting deepfakes through deep learning methods. Convolutional Neural Networks (CNNs) have been mostly applied since they are able to efficiently extract visual features from these manipulated images. Models like MobileNetV2 and ResNet50 are usually used because of their excellent ability to do so, along with being computationally efficient. Nevertheless, it might not be effective to solely rely on deep learning-based techniques for obtaining the optimal results [10] [11].

Recent studies reveal that a combination of deep learning methods with classifiers of machine learning is useful for enhancing detection precision and reliability of the overall framework. Methods of machine learning, such as Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), can be used to make decisions using features obtained from CNN algorithms. Hybrid models tend to increase efficiency, feature exploitation, and generalization power over pure deep learning-based frameworks [12].

Even though advancements have been made to improve deepfake generation and detection systems recently, current facial forgery detectors still experience certain limitations while working with sophisticated manipulations within realistic images. In particular, traditional approaches demonstrate decreased effectiveness in the presence of manipulations performed by means of compression, resizing, filtration, and blurring. Moreover, several current approaches require the use of a deep learning classifier only, meaning the high level of computation costs and inability to work efficiently with diverse manipulated images. Conversely, the fast development of AI-based facial image manipulation has led to the necessity of creating new detection systems, allowing for achieving both effective feature extraction and efficient classification in order to ensure higher levels of performance and detection effectiveness. Thus, there is a significant need for developing an approach for deepfake detection that would involve both deep learning and machine learning algorithms [13], [14].

The reason behind developing a dependable, effective, and generalized model for detecting deepfakes became one of the key areas of study in digital security and forensics. Reliable frameworks are needed to maintain confidence in digital media, develop biometric recognition systems, decrease the dissemination of misinformation, and verify the authenticity of digital data.

The primary goal of this research work is to design a detection model utilizing deep learning and machine learning techniques. In the proposed approach, pre-trained CNNs would be used to extract the features from the images. Machine learning classifiers would be utilized to increase the reliability and performance of the detection algorithm. To achieve our objective, the proposed study was designed following the research questions (RQ):

RQ1: How effectively do transfer learning-based CNN models like MobileNetV2 and ResNet50 help in detecting important facial characteristics to detect deepfake images?

RQ2: Is there any improvement in the detection of deepfake images if we combine ML classifiers such as Random Forest (RF) with deep learning-based facial feature extraction methods?

RQ3: Which is the most suitable and accurate combination of deep learning models and ML classifiers for detecting deepfake facial images?

Some of the key contributions of this research work are as follows:

- A novel framework for detecting deepfakes by implementing machine learning and deep learning approaches has been introduced.
- Deepfakes have been detected by using transfer learning-based CNN architectures such as MobileNetV2 and ResNet50.
- Different machine learning classifiers like RF, SVM, and logistic regression have been utilized to increase classification performance.
- Data augmentation, class imbalance problem, early stopping, and learning rate scheduling have been utilized to enhance generalization performance.
- Experiments were performed based on the ROC curve, confusion matrix, and comparative analysis of accuracy.

This study is significant since it helps in solving one of the biggest issues currently facing digital security: identifying manipulated facial content using Artificial Intelligence. This study presents some contributions to cybersecurity, biometrics, digital forensics, and media validation through the improvement of AI deepfake detection technology. This study is practically applicable to social media monitoring, online identity verification systems, surveillance systems, and misinformation detection, among others. Additionally, the use of deep learning and machine learning in this process gives a precise and effective computational method to help design intelligent security systems in the future.

The rest of the study is structured as follows: Section II identifies the key research gap. In Section III, detailed steps for conducting the proposed study are discussed. Key findings, interpretation, and discussion are explained in Section IV. Section V outcome / future directions are presented.

II. Literature Review

Deepfakes are facial images that are generated using AI. It is a major problem for the security of the web, privacy online, and the authenticity of data present online. Several deep learning architectures can be used to detect DeepFakes. There are several pretrained CNN architectures, such as Inception netv3, Resnet 50, Resnet 101, and DenseNet121, trained using available benchmark datasets such as real and fake face detection using 2k samples, real versus fake face detection using 140k samples, and deepfake and real images using 190k samples. The efficiency of these pretrained CNN models depends on the availability of the number of samples and their resolution. The current pretrained models do not perform well in terms of efficiency; for example, a pretrained model performs with an accuracy of 55% in an imbalanced and limited data set, such as the one with a real and fake data set of 2k samples. In this study, a hybrid DL model combining both the capability of a pretrained model and a GAN with an SE attention mechanism achieves an accuracy of 89% [15]. With the rising cases of counterfeit digital images and increasing adoption of the deep fake technology, there is a need for developing efficient algorithms that can be able to detect counterfeit pictures. This study [16] provides a deep neural network algorithm known as DeepFaceGuard with an objective of detecting counterfeit pictures using a multi-layer architecture in order to ensure low detection time without compromising on accuracy. A pre-trained model of VGG-16 was used to increase visualization, while the principal component analysis technique was used to reduce dimensions. Accuracy and F1-score of 0.83 were some of the measures used to evaluate the proposed model by use of 140,000 image dataset. It was found that DeepFaceGuard had an accuracy of more than 83%. This shows its efficiency in detecting counterfeit images. DeepFaceGuard attained a verification accuracy of 84% in only 184 seconds (8 seconds × 23 epochs). Prior studies involved residual convolutional blocks or the use of an attention mechanism for DF detection; unfortunately, most researchers

consider each of these separately without any emphasis on fairness, out-of-distribution generalization, statistical tests, or artifact-based detection. This study [17] proposes a method named ARC-Net that employs both attention and residual convolutional blocks, along with a popular base, EfficientNet B0, and an attention mechanism. For testing ARC-Net alongside the rest of the four pre-trained DL models, a real image dataset of 500 images from Bangladesh, in addition to a 140k mixed real and fake faces data set, was used. ARC-Net outperformed the traditional and state-of-the-art methods significantly with an accuracy of 99%, a precision of 1.0, a recall of 0.97, and an F1 score of 0.98, obtaining the best possible reliability for DF detection. For evaluating the external validity and generalizability of ARC-Net, two new data sets, the Deepfake dataset and the Deepfake database, have been tested for the model, and it showed consistent and highly reliable results in all experiments conducted, which also include varying scales. Three OOD tests were performed; the first test examined the South Asian dataset, and we observed that including less than one percent of the actual Bangladeshi image data lowered the FPR and enhanced the probability calibration significantly, whereas the other two cross-dataset tests exhibited high transferability. We have also done an ablation test to observe the effect of various components in a model by deleting and changing them, and we have used McNemar's Statistical Test to determine statistical significance between rival classifiers. Furthermore, we have utilized Explainable AI (XAI) approaches such as Grad-CAM and LIME to provide transparency regarding the outcomes, considering that the facial region needs to be prioritized for detection. The fast-paced development of Deep Learning (DL) algorithms has caused the extensive emergence of DeepFake (DF) images in real life, which raises real-life challenges. In the present work, DeepDect is proposed and evaluated within a real-life context. The platform has been developed through a human-centered approach, considering insights (requirements) from both regular users and DF experts. The most effective DL models (ResNet-50 and Random Forest for face-swapping detection, Capsule Forensics v2, and CNN for AI-generated images) have been incorporated into the platform as a detection engine. An Explainable AI (XAI) module has been designed and incorporated into DeepDect for visual (Grad-CAM heatmaps) and textual explanations. Real-life evaluation has been conducted with the participation of 108 users to measure the performance of DeepDect against humans. DeepDect has obtained 81% detection accuracy and outperformed human users in terms of DF detection rate, emphasizing the necessity for such a solution in real-life contexts [18]. With the fast development of face manipulation techniques through Generative Adversarial Networks (GANs) and Stable Diffusion-based approaches, this study [19] presents an investigation into the topic of deepfake content generation. This study examines the possibility of utilizing Vision Transformers (ViTs) for deepfake image detection based on the ability of such networks to extract global features. To determine the efficiency of ViTs for multiclass deepfake image detection was used. The proposed model shows high levels of accuracy, precision, and recall, and the F1 score of 99.90% on a multiclass-prepared dataset. The presented findings show that there is a great potential in ViTs to help make the internet safer from deepfake threats, especially caused by Stable Diffusion and StyleGAN2. The developed model demonstrated better performance than state-of-the-art CNN-based models, i.e., ResNet-50 and VGG-16. Currently, deepfakes have been considered a popular research area because of the considerable improvements made in the field of artificial intelligence (AI). Images produced by the generative AI system are realistic and entirely new. Hence, detecting the fake material using AI is very important. In this study[20], transfer learning was implemented on four models, including VGG16, VGG19, MobileNetV2, and ResNet50V2. The proposed model in this study is the one that is developed based on the pre-trained VGG16 model. In this method, we used the input of the VGG16 model along with the output of two more convolutional and MaxPooling layers for further feature extraction after concatenation, followed by three dense layers. According to the experimental values provided in this study, the proposed model achieves better performance compared to other approaches with 98.4% accuracy rate. It can be concluded that the detection of deepfakes will be feasible in the future through machine learning techniques. Deepfake technology development has outpaced that of video manipulation detection methods, posing some issues in terms of media security and forensic digital analysis. In this case study [21], we implemented a multimodel machine learning approach by implementing traditional as well as modern deep learning classifier models for deepfake video detection. For our case, the experiment made use of a subset of the dataset, which included 401 samples of training and 400 samples of testing data, but with an imbalance in classes in favor of FAKE videos. Our methods involved data preprocessing, EDA, face detection using Haar Cascade Classifiers in OpenCV, and data feature extraction. Models, including Random Forest, Support Vector Machines, and Convolutional Neural Network (CNN), among other models, were trained and tested using the model accuracy, precision, recall, and F1 score as metrics. Our deep learning model performed better than any other model in all metrics, indicating its capability in deepfake video detection. The study above shows how deepfake video detection requires a combination of classical and modern machine learning techniques. With the rise of hyper-realistic fake media generation techniques through deep learning applications, there is a need for detecting the authenticity of images online. In this study [22], an ensemble learning-based technique integrating InceptionV3, VGG16, and Xception architectures was applied to develop a deepfake detection system. The method yields an accuracy of 97.40% with 97.00% sensitivity and 98.00% specificity on a balanced dataset containing 70,000 real and 70,000 fake images. A comparison with ResNet50 and EfficientNet baseline models shows that our model is 5.6% more accurate than ResNet50 and 3.0% more accurate than EfficientNet in terms of AUC. Our findings indicate that the proposed method is promising in addressing the problem of detecting deepfake images. To assess the reliability of the developed model, extensive experiments were performed using a large number of real and synthetic images. In addition, Grad-CAM (Gradient-weighted Class Activation Mapping) was used to interpret the decision of the model. The deepfake videos have been distributed in recent years using several deepfake techniques (i.e., faceswap, face2face, etc.). The aim of this study [23] was to design a model that helps detect deepfake videos. Two hybrid models, consisting of a combination of machine learning and deep learning, were proposed. The first hybrid

model used CNN (convolutional neural network) for feature extraction along with different classifiers in machine learning, namely, SVM (Support Vector Machine), KNN (K-nearest Neighbor), DT (Decision Tree), RF (Random Forest), Logistic Regression, and Naive Bayes. In comparison, the second hybrid model utilized the concept of transfer learning using the VGG16 (Visual Geometry Group) pre-trained model and the same classifiers as the first model. Evaluation of both the hybrid models was done using the video dataset, FaceForensics++, containing four types of deepfake techniques: Deepfake, Faceswap, Face2Face, and Neural texture. Results obtained for both models were found to be promising due to high accuracy scores, proving the effectiveness of the proposed models that can be employed in the construction of a deepfake detector application. The first hybrid model gave maximum accuracy with the SVM classifier for the four deepfake techniques: 0.96, 0.87, 0.90, and 0.64, respectively. The impact of deepfake videos is immense and can range from identity theft to spreading misinformation to national security threats. Hence, there is a dire need to improve the performance of the algorithm used in deepfake detection. In this research [24], various feature extraction methods have been used to leverage the use of deep learning algorithms like Xception and ResNet50 in detecting deepfakes in a video using the DFDC data set. Further, several hybrid models comprising eight were created through the use of classification algorithms like SVM, KNN, MLP, and RF. ResNet50 and RF hybrid model showed the most promising results in terms of the accuracy rate, which stood at 98% and an area under the curve (AUC) of 99.65%. The emergence of manipulation on digital media poses a serious threat in today's society, hence the need to have accurate and reliable means of detection of these fakes. To provide an appropriate response to the problem, the study [25] proposes a novel hybrid structure known as CMNV2, which incorporates MobileNetV2 along with an invented CAFFE block. By simply introducing five extra layers in a pre-trained structure, this architecture is able to withstand even the most complicated situations in the real world to score 99.10% accuracy against 13 other CNN structures. The present study trained and tested the proposed CMNV2 architecture through DNNs, TL, and DL on 5,000 images that were divided into real faces and deepfake faces. When compared to the 13 CNN architectures in terms of some important parameters, the proposed model was superior in terms of accuracy, precision, recall, F1-score, error rate, and efficiency.

III. Methodology

This study presents an integrated method to detect deepfake images based on deep learning and machine learning techniques. The technique relies on the use of pre-trained CNN models to identify key features of the image, which are further used to classify them using conventional machine learning algorithms. The proposed methodology is based on a whole workflow, involving image preparation, dataset splitting, data augmentation, extraction of deep features, classification, and evaluation stages as illustrated in Fig. 1.

1) Data set collection and preprocessing: All the experiments in this study were conducted on the Deepfake Detection Dataset 2026, which was retrieved from the Kaggle portal [26] some samples images are shown in Fig. 2. It is a large dataset for detecting deepfakes, which includes facial images that are specially designed to help analyze deepfakes using machine learning methods. In the given facial image dataset, there are 6,557 images, which can be classified into two categories (2,790 real images and 3,767 fake images). The real images were obtained from Unsplash and include portraits of people of all age groups and genders. The artificial face images were generated by implementing highly sophisticated AI-based image generation algorithms, such as StyleGAN3 another. One of the key aspects that distinguishes this dataset from other commonly used deepfake image datasets is that it comes with comprehensive metadata information, which can be useful for conducting advanced analysis experiments using machine learning. The given dataset includes 17 different attributes that fall into the following categories, represented in Table I.

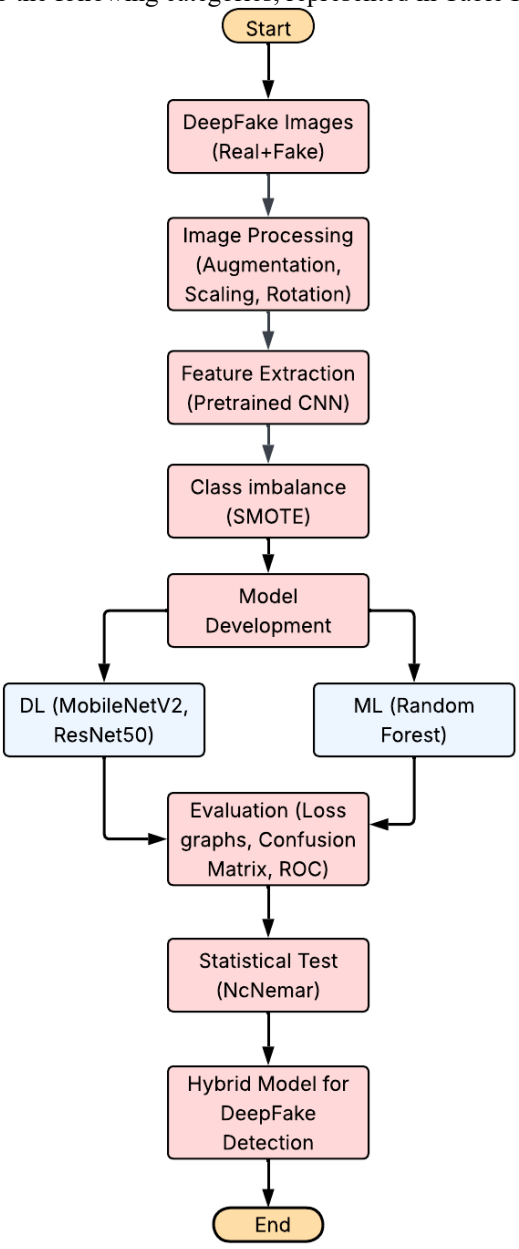


Fig. 1. Overall methodological flow chart of the proposed model.

TABLE I. DATASET CHARACTERISTICS

Dataset Characteristics	Values
Text Attributes	11
Numerical attributes	2
Unique	1
Others meta data	3
Class labels	Real or Fake
Difficulty level	Easy, Medium, Hard
Confidence score	Numeric
Source information	Text
Generation method information	

Only the high-quality faces were used in this dataset.

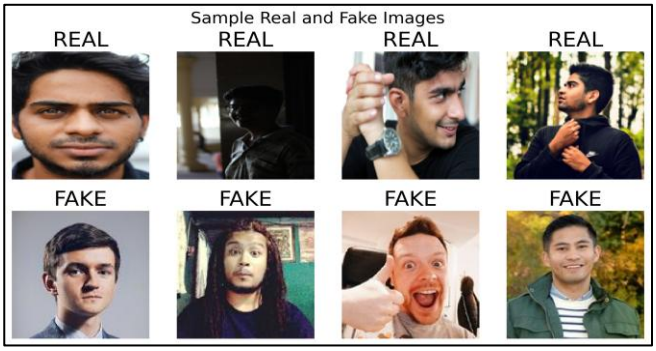


Fig. 2. Some sample images.

2) Image acquisition and preprocessing: All images are downloaded dynamically from URLs using HTTP requests. The preprocessing pipeline is shown in Fig. 3.

Fig. 3. Steps of preprocessing.

This preprocessing standardizes input dimensions and improves neural network convergence stability.

3) Data augmentation: Real-time augmentation of the dataset was done to increase the generalization capabilities of the model and minimize the risk of overfitting by applying augmentation only on the training set via the use of TensorFlow's ImageDataGenerator. Several types of augmentation were done to generate different variations of the face [27]. Fig. 4 shows the data augmentation parameters.

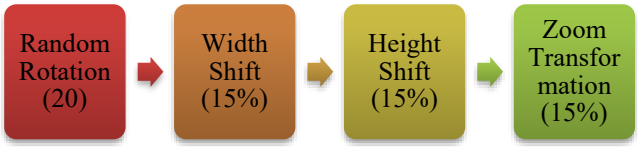


Fig. 4. Data augmentation parameters.

Fig. 4 shows the data augmentation parameters applied to images. This is to ensure that the model performs well and detects deepfake photos regardless of how the person looks in the real world.

4) Feature extraction: Feature extraction is one of the most crucial stages in the proposed system. This research utilizes pretrained convolutional neural networks (CNN) architectures as automatic feature extractors. Unlike handcrafted feature engineering, CNNs learn hierarchical spatial representations directly from facial images [28]. Given an input image be denoted as:

$$X \in \mathbb{R}^{H \times W \times C} \quad (1)$$

The conventional operation with K learnable convolution kernel and the extracted local feature map are represented as $(X * K)$, mathematically defined as:

$$(X * K)(i,j) = \sum_m \sum_n X(i-m, j-n)K(m,n) \quad (2)$$

These layers detect facial textures, blinding artifacts, compression traces, and illumination inconsistencies. The CNN extracts multilevel semantic representation using 3 layers: an early layer that learns low-level patterns using Eq. (3)-(5).

Eq. (3)-(5) represents the hierarchical feature learning conventional neural network, where each layer extracts progressively more abstract representations from the input image. The first layer captures low-level visual patterns learn more complex and discriminative features. The parameters W *and* b denote learnable weights and bias, $\sigma(.)$ denotes the nonlinear activation function.

$$\boldsymbol{F^{(1)} = \sigma\left(W^{(1)} * X + b^{(1)}\right)} \quad (3)$$

$$\boldsymbol{F^{(2)} = \sigma\left(W^{(2)} * F^{(1)} + b^{(2)}\right)} \quad (4)$$

$$\boldsymbol{F^{(3)} = \sigma\left(W^{(3)} * F^{(2)} + b^{(3)}\right)} \quad (5)$$

Furthermore, each convolution layer produces an activation map:

$$F_K = \sigma(W_K * K + b_K) \quad (6)$$

where,

- W_K: convention filters
- b_K: bias term
- σ: activation function

These feature maps are then compressed using global average pooling (GAP). For the feature map F_K :

$$g_K = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} F_K(i,j) \quad (7)$$

Thereby preserving dominant semantic activation.

The final feature representation can be mathematically expressed as:

$$f = [g_1, g_2, \ldots \ldots g_d] \; where \; f \in \mathbb{R}^d \quad (8)$$

This feature representation includes spatial inconsistencies, blind spots, and illumination.

The final feature vector generated through the (GAP) layer was provided as an input to the Random Forest classifier for the classification stage. The final feature vector included high-level features extracted by the network and characterized various aspects related to facial texture changes, lighting changes, compression distortions, and manipulation-related features. In order to make sure that all the features have equal weight in terms of classification, the z-score normalization was performed on the extracted features.

Dimensionality reduction techniques were not used since the GAP layer provided the compact representation of the input features. Thus, the normalized feature vectors were supplied directly into the Random Forest classifier. The classifier was designed using 100 decision trees and the Gini impurity criteria for splitting nodes. The bootstrap method was used to improve the generalization ability of the model. Moreover, the number of features for splitting was selected based on the square root rule. These configurations were chosen since they are commonly used for deep feature classification and have proved their effectiveness in the experimental analysis.

5) Handling class imbalance: There is a minor imbalance of classes, where the number of images that are not real is higher than that of the images that are real. To avoid bias towards the majority class when predicting, the class weights are assigned inversely to their frequencies, as demonstrated in Fig. 5. This was done during the training process, whereby the neural network could be trained to perform well for both classes of images. To solve the problem of unbalanced classes in the training dataset, SMOTE is employed. This approach helps synthesize more artificial samples for the class that is poorly represented by creating new samples based on existing features. Consequently, this improves the ability of the model to generalize from minority examples and eliminates any biases. To preserve the integrity of the evaluation process, SMOTE is not used for the test dataset but only for the training set to achieve the objectives [29].

6) Model development: In this study, 2 pre-trained DL models, such as MobileNetV2 and ResNet50, and 1 ML classifier, such as RF, are used.

a) MobileNetV2: MobileNetV2 was selected for this research due to its compact and efficient architecture that allows for less computational complexity. Depth-wise separable convolutions make it possible for the network to use a minimal amount of parameters and still perform efficient feature extraction.

b) ResNet50: ResNet50 was selected based on its ability to learn features in a deep network thanks to residual connections. Such an approach allows overcoming the problems associated with vanishing gradients and facilitates the process of training deep networks [30]. Models used in this research were pretrained with the ImageNet dataset with the following parameters:

- Include_top=False
- The convolutional part was not trainable during the initial training phase.

c) Random Forest Classifier (RF): RF is an ML classifier used to generate multiple decision trees. Predictions are made through a consensus of those trees, which enhances the accuracy and robustness of predictions. It is efficient in dealing with serious dimensional problems as well as in the identification of nonlinear associations between the behavioral (Fig. 6) [31].

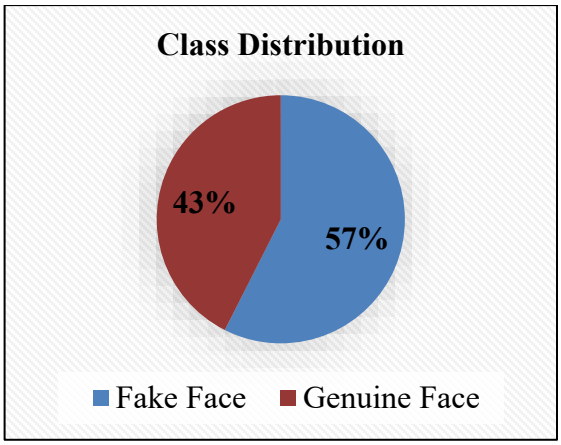


Fig. 5. Class distribution.

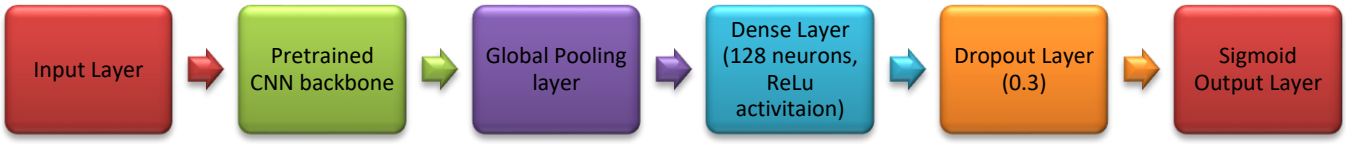


Fig. 6. Deep learning architecture parameters.

7) McNemar test: McNemar's test is a statistical procedure that can be used to make a comparison between two classifiers based on the same set of data. As opposed to testing overall accuracy, it tests the points where the two classifiers disagree. It is thus more appropriate for determining the differences between two classifiers [32].

8) Experimental setup: This experimental setup was designed to evaluate the efficacy of various ML and DL approaches that were utilized to detect real and fake faces. All experiments were conducted within the Python programming framework, utilizing typical libraries for analyzing and modeling data through machine learning techniques. Training was performed for up to 40 epochs with a batch size of 8. In order to make the training more efficient and improve the performance of the model, two callbacks were introduced.

a) Early stopping: The training was terminated automatically if no improvements in the validation loss

occurred within five epochs. This ensured that training stopped at the appropriate point, avoiding any overtraining.

b) Reducing learning rate on plateau: Another callback introduced was related to the adjustment of the learning rate. If the validation loss does not change over three epochs, then the learning rate is adjusted by multiplying with 0.5 [33].

The sigmoid activation function computes the probability of fake classes by:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (9)$$

Moreover, binary cross-entropy was employed as a loss function, which can be mathematically calculated as:

$$l = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ log(\hat{y}_i) + (1-y_i)\, log(1-\hat{y}_i)] \quad (10)$$

The Adam optimizer was used as a gradient optimization because of its adaptive learning-rate capability [34].

To prevent data leakage between subsets, the division of the dataset was made before the beginning of training. All the preprocessing techniques, including SMOTE oversampling and data augmentation, were applied only to the training subset. The validation and test subsets were untouched during the experiments to preserve their properties as representative data. Moreover, the feature extraction process based on the use of pretrained CNN models was performed separately on each subset. Despite the absence of identity, generation source, and near-duplicate relations labeling in the dataset, sample isolation was made during all stages of the experiment. In general, the described experimental configuration allowed obtaining unbiased results regarding the performance of the model on unseen data.

TABLE II. DISTRIBUTION OF SAMPLES

Training Samples	Validation	Testing Samples
2100	297	603

Table II depicts the distribution of samples into training and testing. To ensure consistency and reproducibility, the dataset was organized into a total samples (3000), consisting of both real and manipulated facial images. The dataset was divided into 70% (2100) samples, which were used for training, 297samples (approximately 10%) were used for validation, and 603 samples (approximately 20%) were used for testing. As for ML models, 70% used for training and 30% were used for testing, thereby maintaining consistency and enabling a comparison between the two methodologies. As far as DL methods are concerned, part of the training data set was kept for a validation data set, which helps in monitoring the progress of learning by the model, tuning the hyperparameters of the model, and minimizing the problem of overfitting. Whereas in the case of ML methodology, no separate validation data set was used because the training process involved in ML is much less complex as compared to DL methods.

IV. RESULTS AND DISCUSSION

The objective of this study is to evaluate the performance of various ML and DL classifiers for deepfake detection across multiple measuring standard metrics such as confusion matrix, loss graphs, ROC curve, etc.

TABLE III. PERFORMANCE EVALUATION OF ML MODELS

ML models	Accuracy	Precision	Recall	F1 score
MobileNetV2	98%	90%	91%	90%
MobileNetV2+RF	99%	98%	97%	97%
ResNet50	97%	93%	94%	93%
ResNet+RF	98%	94%	92%	92%

Table III presents the performance of various ML and DL models across all four measuring metrics. The obtained experimental results clearly indicated that hybrid CNNs combined with ML classifiers outperformed CNN models. The best performance was achieved by (MobileNetV2+RF) accuracy (99%), precision (98%), recall (97%), and F1score (97%). RF helps increase the performance of DL and decrease erroneous classifications by integrating feature extraction with accurate classification. Moreover, the ResNet+RF model also showed higher performance rates compared to ResNet50 alone, where accuracy was increased from 97% to 98%. This indicates that hybridization leads to improved feature extraction and discrimination. High performance of MobileNetV2+RF can be attributed to efficient feature extraction of MobileNetV2 through depthwise separable convolutions, which were followed by feature classification via RF. The results have once again proved that hybridizing CNN models with machine learning classifiers yields higher levels of stability and accurate classification compared to deep CNN-based classifiers.

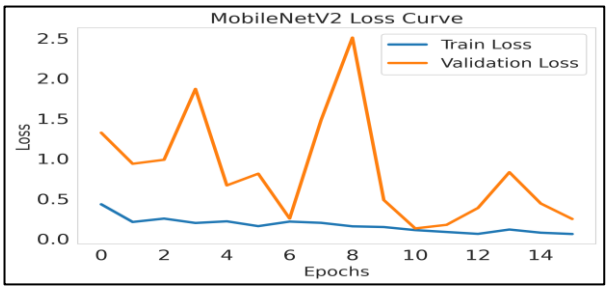


Fig. 7. Loss curves for the MobileNetV model.

Fig. 7 represents the training and validation loss of the MobileNetV2 model over 40 epochs. The decrease of MobileNetV2 training loss appears to be stable, meaning that the network successfully learns from training data and reaches good convergence during optimization. As the training loss consistently decreases, it is reasonable to believe that the network successfully extracts information from the provided training samples. The validation loss demonstrates inconsistent results and experiences sharp increases from time to time. This indicates the problem of overfitting when the neural network performs well on training data but has poor generalization capabilities when processing unknown samples. Altogether, even though the network successfully learns from training data, it requires further refinement to improve its generalization capabilities.

MobileNetV2 Validation Loss Plot demonstrates some fluctuations throughout the training process and even some

spikes. Such behavior might be a hint at a minor case of overfitting, but despite that, the overall plot tends to converge, which implies that learning is taking place. Fluctuations may be associated with the dataset variation and a relatively small number of records in the validation dataset. To improve the training stability, future improvements might include more aggressive regularization, changing the dropout probability, or using data augmentation.

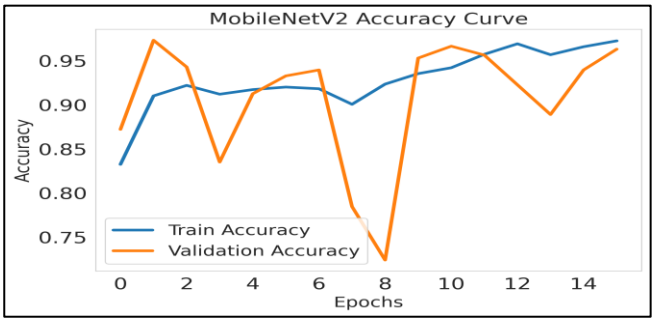


Fig. 8. Model accuracy curve.

Fig. 8 depicts the accuracy graph of the MobileNetV2 algorithm; the accuracy of the learning process by the network is good. With an increase in the epoch numbers, the training accuracy of the algorithm has been seen to increase. The validation accuracy increased along with the training accuracy; however, there have been slight differences in certain epochs due to differences in the validation data. By the last epochs of training, the validation accuracy and training accuracy are both seen to be around 96–97%. This suggests that the algorithm performed well and that there was no overfitting.

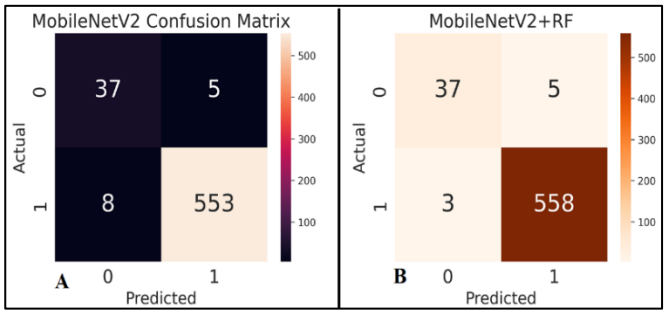


Fig. 9. Confusion matrix for MobileNetV2.

Fig. 9 shows the confusion matrix of MobileNetV2 (A) and MobileNetV2+RF Classifier (B). In Fig. 9(A), the total number of samples accurately classified as negative is 37, and the positive samples are 553. Also, 5 positive samples were wrongly classified as negative, and 8 positive samples were wrongly classified as negative. After using the RF Classifier with MobileNetV2, the correct positive samples classified were 558, and there were only 3 samples classified as negative from the positive class. There was no difference in correctly classified negative and false positive samples, which are 37 and 5, respectively. This implies that the combined model had high accuracy in terms of recall value for positive samples. The high effectiveness is due to the efficient use of the deep learning model by the RF Classifier. The hybrid model used ensemble learning to improve class separability and decrease the false positives in the positive class.

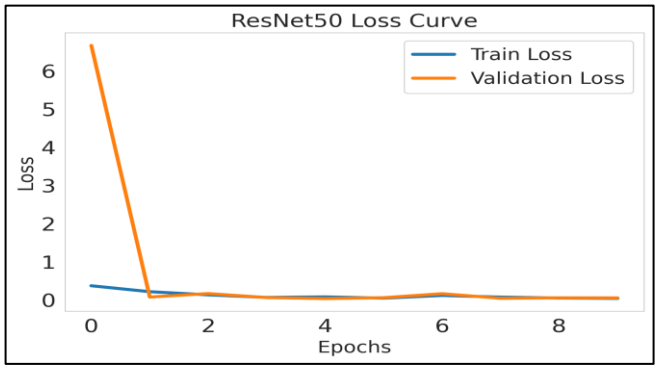


Fig. 10. Train and validation loss curves.

ResNet50 exhibits a smooth and constant learning curve throughout the entire training process, as shown in Fig. 10. Initially, the validation loss is high due to the fact that the model is trained with random weights in the early stages of training. However, following the first epoch, there is a sharp decline in the validation loss such that it is close to the training loss. The two lines are low for the rest of the training period and move almost in parallel without significant differences. Thus, the model successfully learns the critical elements of the data without overfitting. Moreover, the small gap between training and validation losses implies good generalization.

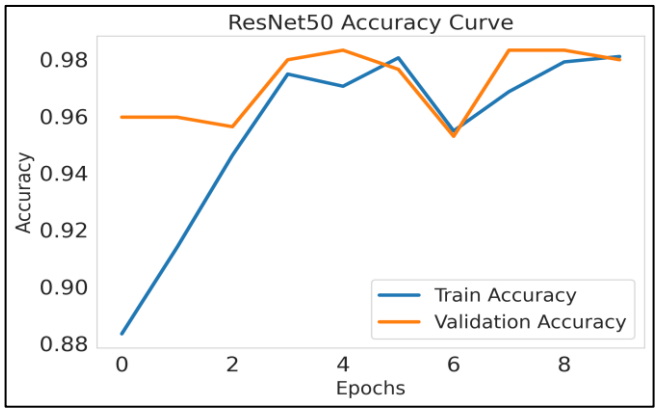


Fig. 11. Accuracy curves for ResNet50.

From Fig. 11, it can be concluded that the performance accuracy of the ResNet50 model is both consistent and accurate over 10 training epochs. The training and validation accuracy of the model increases gradually, and it eventually becomes almost 98%. Initially, the training accuracy starts at a point of 88%, and this implies that the model is learning important characteristics. The validation accuracy is fairly high and has small variations, meaning that the model does not suffer from much overfitting, and it has excellent generalization capabilities. This is evident since the training and validation graphs remain close for all epochs, meaning that the model does not experience problems related to overfitting and underfitting. There is no wide gap between the two lines, which implies that the ResNet50 model works well since it allows easy learning of deep networks. The residual learning technique makes gradient flow easy, and consequently, efficient learning becomes possible.

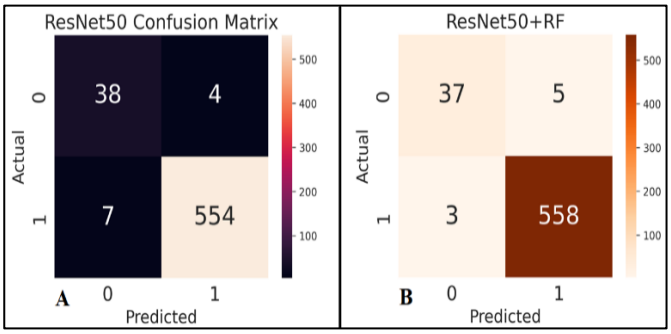


Fig. 12. Confusion matrices for ResNet50.

Fig. 12 shows the confusion matrices for standalone ResNet50 and ResNet50+RF models. Both the ResNet50 and ResNet50+RF models have an extremely high efficiency level since the majority of samples were predicted accurately. Specifically, the model of ResNet50 managed to correctly predict 38 samples of class 0 and 554 samples of class 1, misclassifying only several other samples. However, the combination of ResNet50 and RF increases the accuracy of prediction, lowering the false negatives to 3 (previously 7) and raising the number of correct predictions for class 1 from 554 to 558. It means that adding the Random Forest model to the ResNet50 algorithm significantly contributes to separating the classes correctly. Thus, the hybrid approach ensures more accurate and stable predictions than ResNet50 alone.

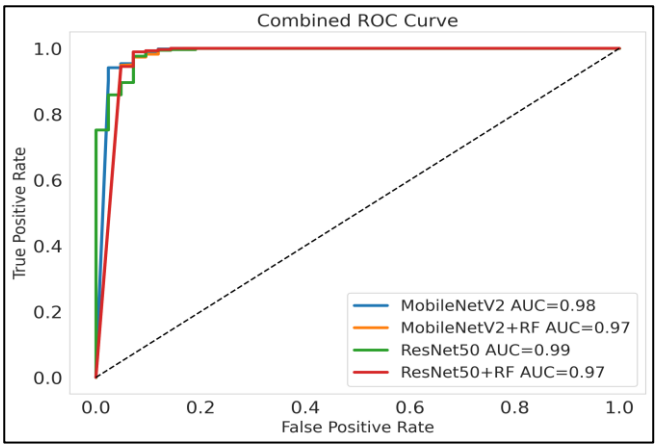


Fig. 13. ROC analysis of models.

Fig. 13 presents the ROC curve analysis results of all models analyzed. From these, the ResNet50 model managed to obtain the highest AUC value of 0.99, reflecting its superior performance in distinguishing classes. The MobileNetV2, with an AUC value of 0.98, while the RF with MobileNetV2(MobileNetV2+RF) model and the RF with ResNet50 (ResNet5+RF) model achieved an AUC value of 0.97. The findings show that the accuracy obtained by these models is quite high, and almost negligible errors were made. From these good AUC values, we can infer that our proposed deep learning models have worked well in differentiating between positive and negative classes. The ResNet50 model has performed better mainly due to its architecture, which enables efficient learning of deeper features in its structure.

TABLE IV. STATISTICAL TEST

Models	Statistic	P-Value
MobileNetV2 vs MobileNetV2+RF	3.2	0.073
MobileNetV2 vs ResNet50	0.05	0.823
MobileNetV2 vs ResNet50+RF	1.067	0.301

Table IV shows the statistical significance results of the comparative analysis among the evaluated models. The findings revealed that there is no statistically significant difference between, as all p-values are greater than the significance threshold of 0.05. Specifically, the comparison between MobileNetV2 and MobileNetV2+RF yields a test statistic of 3.2 with a p-value of 0.073, suggesting marginal improvement, this improvement is not significant. Moreover, it is possible to note that for MobileNetV2 and ResNet50, the p-value equals 0.823, while for MobileNetV2 and ResNet50+RF, it equals 0.301. In both cases, these models have almost similar classification performance levels. Thus, the analysis shows that all the methods provide rather good and quite stable results.

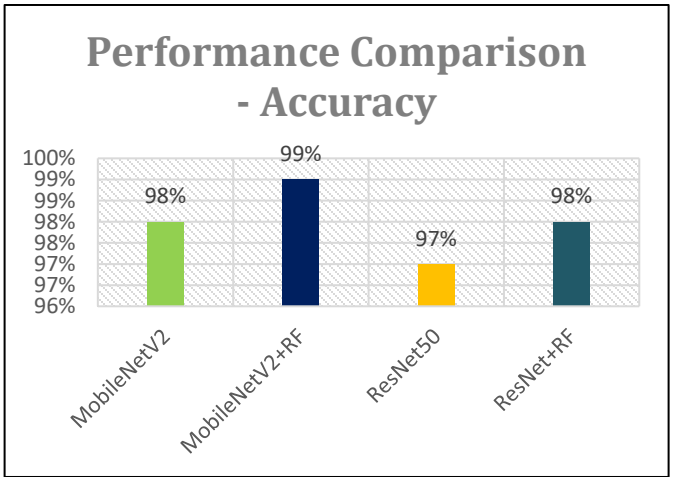


Fig. 14. Comparison of performance of ML/DL models.

In Fig. 14, the comparison of performance in terms of the accuracy metric is shown. The (MobileNetV2 + RF) achieved the highest accuracy rate (99%), attributed to its ability to classify well compared to other models. The model accuracy of MobileNetV2 and (ResNet50 + RF) was same (98%). The difference between their model accuracies is insignificant since they both differ by only 1%. Therefore, the inclusion of the RF algorithm in the deep features helps increase performance stability in classification. The accuracy of ResNet50 on its own was 97% as opposed to other models analyzed in this project. This indicates that the combination of hybrid and deep learning models provides excellent results in classification due to the high accuracy achieved.

V. DISCUSSION

The results of the experiment prove the statement that hybrid models of DL and ML provide better outcomes in terms of deepfake detection accuracy than single-layered CNN architectures, as illustrated in Table II and Fig 14. In particular, MobileNetV2 + RF appears to be the best-performing model that provides the highest accuracy rates among all other models

due to the positive influence of Random Forest on CNN-based image processing. Both MobileNetV2 and ResNet50 models perform effectively without any additional machine-learning model being applied, as illustrated in Table II. The accuracy of both models equals 98% and 97%, respectively. With the help of random forest, the performance of ResNet50 increases by one percent, reaching 98%. All these data suggest the effectiveness of hybrid models, as it was mentioned above. Fig. 14 proves the effectiveness of mobileNetV2 + RF once again.

Though MobileNetV2+RF had the best numerical results compared to other models concerning most metrics, according to McNemar's paired statistical test, there is no significant difference between the performance of the models ($p > 0.05$). It simply means that the performance difference is not statistically significant. Despite the slightly higher accuracy and better stability in the results of the hybrid model, we still cannot prove that there is any practical benefit of such a network over independent convolutional neural networks statistically. Consequently, it should be understood that MobileNetV2+RF only demonstrates better trends in performance. However, considering the improvement in different evaluation criteria, we can state that hybrid CNN architectures can be used in practice to detect deepfakes.

Furthermore, ROC results illustrated in Fig. 13 confirm these findings. ResNet50 has achieved the best AUC score, equal to 0.99, whereas MobileNetV2 is a close second, having an AUC score of 0.98. Hybrid models have obtained an AUC score of 0.97. These findings suggest that all three models are good classifiers with slight differences in their ability to separate classes. Additional confusion matrix results from Fig. 9 and 12 clearly demonstrate that hybrid models effectively decrease the number of false negatives and classify images correctly even in challenging conditions, especially when the Random Forest model is applied to CNN-based feature descriptors. Stable training behavior depicted in Fig. 7, 8, 10, and 11, as well as the absence of any overfitting problems observed with ResNet50, proves that hybrid models can be considered good general classifiers.

In conclusion, it can be stated that Tables II and III, as well as Fig. 7-14 illustrate the superior performance, reliability, and robustness of hybrid models when compared to standalone CNN-based classification methods.

RQ1: How effectively do transfer learning-based CNN models like MobileNetV2 and ResNet50 help in detecting important facial characteristics to detect deepfake images?

According to the results depicted in Table II and Fig.10 and 11, both architectures, namely, MobileNetV2 and ResNet50, demonstrate excellent ability to detect discriminative facial features that could help distinguish deepfakes. The accuracy rate is high for both cases, with MobileNetV2 achieving 98% and ResNet50 reaching 97%. The learning process can be clearly traced with the help of the training and validation curves depicted in Fig. 11 (ResNet50) and Fig. 8 (MobileNetV2), where one can notice the stable and consistent behavior of the curves with a slight gap between the performance of the models during training and validation procedures. Moreover, the ROC curves plotted in Fig. 13 prove the results obtained earlier because both models show high AUC values of 0.98 and 0.99. In other words, they demonstrate their ability to differentiate real images and images that have undergone manipulations.

RQ2: Is there any improvement in the detection of deepfake images if we combine ML classifiers, such as RF, with deep learning-based facial feature extraction methods?

Indeed, the fusion of the RF classifier with feature extractors based on CNNs provides a remarkable enhancement in classification performance according to Table II and Fig. 14. The hybrid architectures, namely MobileNetV2+RF and ResNet50+RF, always surpass the performance metrics provided by the plain CNN classifiers. Specifically, MobileNetV2+RF achieves the highest value of accuracy (99%), while ResNet50+RF improves the accuracy score from 97% to 98%. One can assume that RF assists in improving the decision-making process based on the deeper feature representation obtained via the CNN architecture. Furthermore, Fig. 9 and 12 illustrate that there are fewer cases of false negatives once the Random Forest technique is applied to classification, particularly when considering the performance metrics of MobileNetV2+RF. Thus, it becomes easier to distinguish between the real and manipulated images. As a result, the proposed hybrid approach demonstrates its effectiveness since a fusion of classical machine learning and CNNs helps improve the classification process significantly.

RQ3: Which is the most suitable and accurate combination of deep learning models and ML classifiers for detecting deepfake facial images?

Based on the results obtained from the comparative analysis shown in Table II, as well as the results shown in Fig. 14 and the ROC curves in Fig. 13, it can be seen that the combination of the MobileNetV2+RF model produces the best results for deepfake video identification. This model achieves 99% accuracy, 98% precision, 97% recall, and an impressive 97% F1-score. ResNet50, on the other hand, achieves a higher area under the curve of 0.99 in Fig. 13; however, it falls short of MobileNetV2+RF in terms of performance in other evaluation parameters. The statistics provided in Table III indicate no statistically significant differences among all the models used for testing ($p > 0.05$). This means that despite some deviations, there is no strong statistical evidence regarding the differences between them. Nevertheless, MobileNetV2+RF remains the best option in terms of reliability.

Across all comparisons in this study, the p-values were consistently above 0.05. This implies that despite the small improvements observed in the hybrid models' accuracy, there is no statistical significance in the differences. Generally, the models exhibit relatively even behavior in terms of classification ability.

The proposed framework shows better results for all used metrics, but the experiments were carried out over a single dataset, which makes it difficult to be sure about how the results are generalizable to any other conditions. The absence of cross-dataset experiments and k-fold validation on various deepfake benchmarks makes it difficult to evaluate the robustness of the model to manipulation types that could be applied to deepfakes in practice. Although the presented results indicate good

classification accuracy, more experiments need to be done to confirm the robustness of the model to various datasets.

1) Practical implications: The proposed approach holds practical relevance when applied to detect deepfakes. It can be applied in social networks to automatically detect manipulated images that contain fake information. The efficiency and reliability of the solution mean it could be used in cybersecurity to prevent identity theft and detect fraud. In particular, since the MobileNetV2 is rather lightweight, the solution can be applied even on devices with low computational power, like mobile phones or edge devices. Utilizing the MobileNetV2 as a feature extractor allows making the architecture of the neural network more compact and lowering the computational cost associated with running it. Additionally, the decision not to employ the full-fledged deep learning system in favor of using a Random Forest classifier significantly decreases the complexity of training the model. Additionally, due to the combination of deep learning algorithms with traditional machine learning, the approach can be scaled and used not only for images but also for other data types. The aspects like training time, inference speed, the size of the trained model, FLOPS, and memory requirements have not been evaluated. While these aspects have been considered during design, there was no formal evaluation done. Computational cost is a question that needs to be investigated in the future.

2) Limitations: Although the findings have shown high effectiveness, some limitations still exist in the experiment. First, the dataset, although diverse, is still limited in size when compared to the vast array of deepfake samples in practical situations, thus affecting the model's ability to accurately classify unknown samples that involve more complicated manipulation. Secondly, the proposed method utilizes CNN models trained by other researchers, and therefore, might not necessarily be adaptable enough for new types of deep fakes.

Although the proposed framework was successful in deepfake image recognition, the current study does not include any ablation analysis, through which the specific impact of each component would be examined. Therefore, it is difficult to understand how much the data augmentation, class balancing through the SMOTE technique, CNN-based feature extraction, and Random Forest classification have affected the outcome of the research.

Further study of these components is very important for the future. Through the process of excluding, replacing, and changing one element after another, it will be easier to see how each of the components affects the performance of the proposed framework.

Recent studies suggest that Vision Transformers (ViTs) achieve impressive results in detecting deepfakes, mainly due to the ability to encode long-range dependencies between different parts of an image. The ViT-based approaches have been mentioned in the literature review in order to represent recent advances in the field, but no comparisons with these approaches have been included in this experiment.

Conversely, the proposed study chosen to tackle the problem of detecting deepfakes is a pipeline, consisting of a pretrained CNN-based feature extractor and a Random Forest classifier. Such an architecture has lower computational complexity and is more suitable for working with small imbalanced datasets. For this reason, a comparison with ViT-based approaches has not been performed, as the latter ones have higher demands on the computational resources and a training procedure.

The future direction of this research includes the implementation of ViT-based architectures in order to compare their performance with the state-of-the-art approaches to deepfake detection.

VI. Conclusion

In this study, we presented a hybrid system that uses deep learning and machine learning for recognizing deepfake images, where deep learning with a CNN feature extractor and machine learning with Random Forest were used to improve the performance of each algorithm on its own. Experimental results revealed that MobileNetV2 + RF provided the highest value of accuracy, equal to 99%, whereas ResNet50 + RF and MobileNetV2 were among those with the best performance results. A statistical analysis of the obtained results revealed that there is no significant difference between the performances of various algorithms, implying a stable behavior of all models. Therefore, our work increases both accuracy and reliability.

Future research in this field could focus on increasing the effectiveness of transfer learning for generalizing to novel and unseen deepfake datasets that have been produced by using state-of-the-art deepfake generation techniques. It can also be interesting to explore the applicability of lightweight transformers since they can assist in learning and representation efficiency improvements. Evaluating these methods using other kinds of data, including video streams and live applications, could broaden the scope of use of the proposed system. Furthermore, efficiency improvement through model compression and pruning would enable real-time processing, especially on edge and mobile devices where computational power is limited.

Data Availability Statement: The dataset used in this study for the detection of deepfake images is publicly available at https://www.kaggle.com/code/engrgul/backdoor-ai

References

[1] P. Rathi, “Unveiling the Future: An Introduction to AI-Powered Multimedia Data Processing,” in AI Techniques for Multimedia Data Processing, IGI Global Scientific Publishing, 2025, pp. 247–268. doi: 10.4018/979-8-3693-2935-1.ch009.

[2] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, “Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning,” Information, vol. 16, no. 4, p. 270, Apr. 2025, doi: 10.3390/info16040270.

[3] C. Zhao, C. Wang, Z. Song, G. Hu, L. Wang, and D. Miao, “Multi-definition Deepfake detection via semantics reduction and cross-domain training,” Pattern Recognition, vol. 163, p. 111469, Jul. 2025, doi: 10.1016/j.patcog.2025.111469.

[4] S. Alanazi and S. Asif, “Exploring deepfake technology: creation, consequences and countermeasures,” Hum.-Intell. Syst. Integr., vol. 6, no. 1, pp. 49–60, Dec. 2024, doi: 10.1007/s42454-024-00054-8.

[5] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, “LightFakeDetect: A Lightweight Model for Deepfake

Detection in Videos That Focuses on Facial Regions," Mathematics, vol. 13, no. 19, p. 3088, Jan. 2025, doi: 10.3390/math13193088.

[6] N. Negi, D. Kaushal, S. B. Walia, and S. Fulara, "The Exposure of Deepfake Technology on Social Media and its Effects on Society," in Mastering Deepfake Technology: Strategies for Ethical Management and Security, River Publishers, 2025.

[7] I. Sagheer, "THE ROLE OF DEEPFAKE POLITICAL VIDEO IN SHAPING IDEOLOGICAL DISCOURSE ON INSTAGRAM," Journal of Applied Linguistics and TESOL (JALT), vol. 8, no. 3, pp. 356–370, Jul. 2025, doi: 10.63878/jalt968.

[8] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[9] J. Jheelan and S. Pudaruth, "Using Deep Learning to Identify Deepfakes Created Using Generative Adversarial Networks," Computers, vol. 14, no. 2, p. 60, Feb. 2025, doi: 10.3390/computers14020060.

[10] O. A. Shaaban and R. Yildirim, "Audio Deepfake Detection Using Deep Learning," Engineering Reports, vol. 7, no. 3, p. e70087, 2025, doi: 10.1002/eng2.70087.

[11] Dr. C.V. Raman University, India, A. Pandey, A. K. Tiwari, and Dr. C.V. Raman University, India, "ADVANCING FACE SPOOFING DETECTION WITH LBP, PCA, AND SVM: A ROBUST AI SECURITY APPROACH," IJIVP, vol. 15, no. 4, pp. 3589–3595, May 2025, doi: 10.21917/ijivp.2025.0508.

[12] A. M. Alashjaee, "Machine Learning Approach for Fake Image Forensics," University of Bisha Journal for Basic and Applied Sciences, vol. 1, no. 1, Sep. 2025, doi: 10.65073/3122-3508.1005.

[13] L. A. Passos et al., "A review of deep learning-based approaches for deepfake content detection," Expert Systems, vol. 41, no. 8, p. e13570, 2024, doi: 10.1111/exsy.13570.

[14] S. S. Patil, A. M, P. M, and P. P. C, "AI/ML-Based Solution for Detecting Face-Swap Deepfakes," in 2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T), Jun. 2025, pp. 1–7. doi: 10.1109/ICETI4T63625.2025.11132146.

[15] U. M and J. Bhimavarapu, "Design and development of hybrid deep learning technique for deep fake identification: in limited dataset perspective," Eng. Res. Express, vol. 8, no. 2, p. 025221, Jan. 2026, doi: 10.1088/2631-8695/ae3664.

[16] N. S. Sagheer, Z. Thamer, A. A. Abdulmunem, H. Thamer, O. Ata, and E. Mohammed, "Deepfaceguard for Real and Fake Face Classification," Apr. 2026, doi: 10.12785/ijcds/1571166529.

[17] M. S. Reza, F. Elias, M. I. Mahmud, and N. Ahmed, "Attention and residual mechanism-based CNN architecture (ARC-Net) with enhanced fairness generalization for deepfake facial image detection," PLOS ONE, vol. 21, no. 4, p. e0340099, Apr. 2026, doi: 10.1371/journal.pone.0340099.

[18] F. Castro, V. Gattulli, D. Impedovo, and A. Monaco, "DeepDect: an explainable AI platform for face swapping and face generation DeepFake detection," Multimed Tools Appl, vol. 85, no. 5, p. 500, May 2026, doi: 10.1007/s11042-026-21672-1.

[19] M. A. Arshed, S. Mumtaz, M. Ibrahim, C. Dewi, M. Tanveer, and S. Ahmed, "Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model," Computers, vol. 13, no. 1, p. 31, Jan. 2024, doi: 10.3390/computers13010031.

[20] N. A. Almalki, M. Ragab, and F. Kateb, "Leveraging AI for Sustainable Deepfake Human Face Detection Using Transfer Learning Technique," in 2025 22nd International Learning and Technology Conference (L&T), Jan. 2025, pp. 228–232. doi: 10.1109/LT64002.2025.10940123.

[21] M. M. Rahman, M. M. R. Computer, N. Das, M. S. Uddin, R. S. Shammah, and M. M. Rahman, "An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach," in 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS), May 2025, pp. 1–8. doi: 10.1109/ICBATS66542.2025.11258249.

[22] M. B. Jabra, O. Cheikhrouhou, and A. BenAmor, "Deepfake detection through ensemble learning," Multimed Tools Appl, vol. 84, no. 36, pp. 45417–45452, Nov. 2025, doi: 10.1007/s11042-025-20932-w.

[23] N. K. Alzurfi and M. S. Altaei, "A Hybrid Model of Deep Learning and Machine Learning Methods to Detect Deepfake Videos," Iraqi Journal of Science, pp. 736–750, Feb. 2025, doi: 10.24996/ijs.2025.66.2.17.

[24] A. Koçak, M. Alkan, and S. M. Arıkan, "Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach," Politeknik Dergisi, vol. 28, no. 3, pp. 957–968, Jun. 2025, doi: 10.2339/politeknik.1523983.

[25] B. A. Kumar et al., "Hybrid CMNV2: DeepFake faces classification and recognition using deep learning methods," Results in Engineering, vol. 28, p. 107513, Dec. 2025, doi: 10.1016/j.rineng.2025.107513.

[26] "Deepfake Detection Challenge." Accessed: May 22, 2026. [Online]. Available: https://kaggle.com/deepfake-detection-challenge

[27] C. Fan et al., "A novel imbalance fault diagnosis method based on data augmentation and hybrid deep learning models," Structural Health Monitoring, vol. 25, no. 1, pp. 281–302, Jan. 2026, doi: 10.1177/14759217241291143.

[28] Y. Li, H.-N. Wu, H.-Z. Cheng, D.-S. Guo, and R.-P. Chen, "Application of an innovative BMask R-CNN-DAT enhanced with feature extraction and boundary awareness in tunnel leakage detection," Tunnelling and Underground Space Technology, vol. 167, p. 107074, Jan. 2026, doi: 10.1016/j.tust.2025.107074.

[29] C. Tian and R. Li, "Predicting Bond Defaults in China: A Double-Ensemble Model Leveraging SMOTE for Class Imbalance," Big Data and Cognitive Computing, vol. 10, no. 3, p. 81, Mar. 2026, doi: 10.3390/bdcc10030081.

[30] S. Riyadi, F. A. Abidin, and N. Audita, "Comparison of ResNet50V2 and MobileNetV2 Models in Building Architectural Style Classification," in 2024 International Conference on Intelligent Systems and Computer Vision (ISCV), May 2024, pp. 1–8. doi: 10.1109/ISCV60512.2024.10620099.

[31] B. Guo et al., "Exploring the effect of the architecture morphology on urban ventilation at block scale using CFD-GIS and random forest combined method," Sustainable Cities and Society, vol. 122, p. 106241, Mar. 2025, doi: 10.1016/j.scs.2025.106241.

[32] Z. O. Aremu and J. O. Akanni, "Comparative Analysis of Mcnemar's Test and Liddell's Exact Test for Assessing Marginal Homogeneity," KolaDaisi University Journal of Applied Sciences, vol. 2, pp. 87–102, Aug. 2025, doi: 10.5281/zenodo.17438474.

[33] M. N. Razali, N. Arbaiy, P.-C. Lin, and S. Ismail, "Optimizing Multiclass Classification Using Convolutional Neural Networks with Class Weights and Early Stopping for Imbalanced Datasets," Electronics, vol. 14, no. 4, p. 705, Jan. 2025, doi: 10.3390/electronics14040705.

[34] A. Ahmed et al., "Efficient Melanoma Detection Using Pixel Intensity-Based Masking and Intensity-Weighted Binary Cross-Entropy," International Journal of Imaging Systems and Technology, vol. 35, no. 5, p. e70179, 2025, doi: 10.1002/ima.70179.

A Bibliometric Mapping of Transformer-Based Misinformation Detection: Trends and Gaps

Borhan Ab Rahman, Mohd Zakree Ahmad Nazri, Mohd Ridzwan Yaakub
Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia

***Abstract*—The rapid spread of misinformation on social platforms has intensified research on automated detection, with Transformer-based architectures becoming a primary technical foundation. However, the volume and diversity of publications make it difficult to track the field's evolution. It is also challenging to identify the contributors, influential studies, and remaining gaps that shape its knowledge base. This study presents a bibliometric mapping of Transformer-based research on misinformation detection, based on Scopus records retrieved on 3rd March 2026 (n = 3,637). Performance indicators and science mapping analysis are used to profile publication growth, subject-area distribution, leading countries, institutions, and authors, citation impact, and thematic structure. Results show that conference proceedings (47.18%) and journals (41.93%) dominate dissemination, reflecting both rapid-cycle publishing and archival consolidation. Computer Science accounts for most publications (84.99%), while Social Sciences and Medicine contribute non-trivial shares, indicating applied and societal engagement. China, the United States, and India lead in national output, and institutional productivity is concentrated in a small set of research hubs. Citation indicators (h-index = 82; g-index = 154) suggest a broad influence with a distinct, highly cited core. Keyword mapping confirms a stable emphasis on NLP-driven detection in social media and shows recent growth in multimodal approaches and generative-AI-related topics. The findings provide an evidence-based overview of current trends and emerging directions, helping researchers position future work and prioritize underexplored problems.**

Keywords*—Bibliometric analysis; misinformation detection; Transformer architectures; keyword co-occurrence; citation analysis; Scopus

I. Introduction

Over the past decade, decentralized social media platforms have increased the reach and velocity of online communication. The same properties also increase exposure to misinformation by lowering moderation barriers and enabling rapid cross-community spread, including across languages. The consequences extend beyond online discourse and affect democratic decision-making, public health behavior, and social cohesion. These risks have intensified demand for automated detection systems that can support early identification and timely mitigation [1], [2]. In this study, misinformation is used as an umbrella term, while fake news and disinformation are treated as closely related labels reported in the computational literature.

Research in Natural Language Processing (NLP) has, therefore, moved from feature engineering and classical classifiers toward Transformer-based architectures for contextual representation learning. Early misinformation detection pipelines were typically built around handcrafted lexical or syntactic features with classical classifiers [3]. These approaches remain computationally efficient, but they often degrade under implicit claims, context-dependent semantics, and distribution shifts across topics, platforms or time [4]. Transformer architectures have changed the technical baseline by learning contextual representations through attention mechanisms [5]. Models such as BERT, RoBERTa and more recently, large language models (LLMs) have become standard choices for misinformation detection because they achieve strong performance on benchmark datasets and transfer effectively across varied text genres [6], [7]. The research frontier has also moved beyond text-only classification toward multimodal detection, explainability and robustness under real-world conditions [8], [9].

Despite the volume of technical work, the field lacks a precise macro-level account of Transformer-based misinformation detection as a distinct research stream. Existing reviews often treat misinformation or disinformation as a broad phenomenon and prioritize diffusion patterns, societal implications or general computational approaches [10], [11]. This leaves several important questions unresolved. It remains difficult to identify which venues and publications define the core knowledge base, how collaboration and influence are structured and which themes are consolidating versus emerging, especially given the recent surge of LLM-related research and the growing reliance on multimodal content. Without this structured view, researchers risk duplicating effort, selecting outdated baselines and evaluating systems in settings that do not reflect current threat models [12].

This study addresses that gap through a focused bibliometric mapping of Transformer-based architectures for misinformation detection. Using Scopus records retrieved on 3rd March 2026 and science mapping techniques, the analysis integrates performance indicators with network-based mapping of collaboration patterns, citation impact and thematic evolution [13], [14]. The study also synthesizes research gaps that follow directly from observed publication and topic trends, supporting clearer prioritization of future work.

A. Related Work

Bibliometric and survey studies have examined misinformation and disinformation research from broad perspectives, including publication growth, influential sources and topic structures. For example, Wang et al. [11] provide a bibliometric review of disinformation research, while Xu et al. [10] map evolutionary themes in misinformation dissemination

on social media. Barve and Mulay [13] survey incremental learning for false-information text classification, reflecting continued interest in efficiency and continual adaptation.

These works provide useful coverage of the wider misinformation landscape, but they do not isolate Transformer-based architectures for misinformation detection as a distinct technical stream. Recent reviews often aggregate heterogeneous model families and settings, which can blur how Transformer-centric methods, multimodal pipelines and LLM-related topics are reshaping evaluation practices and research priorities [14]. A focused mapping that integrates performance indicators with science mapping is therefore needed to clarify the intellectual base, contributor structure and thematic evolution specific to Transformer-based misinformation detection.

This study addresses the following research questions. (RQ1): How has the publication output and citation impact of Transformer-based research on misinformation detection evolved? RQ2: Which articles and sources have the greatest impact in this domain? RQ3: Which countries and institutions contribute the most to the literature? RQ4: Who are the most productive authors and what collaboration patterns dominate the field? RQ5: What thematic clusters dominate the domain and how have they evolved based on keyword co-occurrence and overlay analysis? RQ6: How is the literature distributed across subject areas and what does this distribution suggest about interdisciplinarity? RQ7: What research gaps and future directions emerge from the observed trends in methods, applications and evaluation practices?

Section II details the Scopus data source and search strategy, the inclusion workflow and the procedures used for performance profiling and science mapping. Section III presents the results aligned with the research questions and Section IV summarizes contributions and future directions.

II. Materials and Methods

The data source, search strategy, inclusion criteria and analysis workflow used for this bibliometric study are described in this section. The differences between prior bibliometric reviews and this study are summarized in Table I and the retrieval and screening workflow is shown in Fig. 1.

TABLE I. Comparison of Previous Studies of Misinformation Detection and Our Study

Elements of comparison	[13]	[11]	[10]	Our study
Coverage year	2000–2019	2002–2021	2013–2023	2017 – 2026 (as on 03 March 2026)
Database	Scopus and Web of Science	Web of Science (SCI-E and SSCI)	Web of Science Core Database	Scopus
Keywords	Master keyword "incremental learning" with primary keywords including "false information" OR "misinformation" OR "text classification" etc. and secondary exclusion terms	disinformation OR misinformation OR "fake news" OR "infodemic" OR "information pollut*" OR deepfake* OR "rumor propagation"	TS query combining misinformation terms with social media/social network and diffusion verbs	Misinformation terms combined with Transformer terms
Focus of study	Maps incremental learning for text classification in false information detection and reports publication statistics, collaboration, network and citation analysis plus techniques used	Bibliometric analysis of disinformation research, identifying influential works and central topics using co-citation and keyword co-occurrence	Topic evolution of misinformation dissemination on social media; identifies hotspots and trends using a staged framework (community detection, TOPSIS, visualization)	Bibliometric mapping of Transformer-based architectures for misinformation detection: intellectual structure, dominant clusters, trend trajectory and research gaps
Methodology	Bibliometric survey with publication statistics, collaboration degree, collaboration visualization and citation analysis	Bibliometric analysis: performance analysis and science mapping	Scoping review framework combined with bibliometric analysis	Bibliometric analysis: performance analysis + science mapping
Protocol/Tools	Not explicitly tool-focused in the study, but the objectives include collaboration visualization and citation analysis.	Derwent Data Analyzer (DDA 10) for cleaning, mining and visualization	ITGinsight for preprocessing; integrates community detection, AHP, TOPSIS and visualization	VOSviewer v1.6.20 (mapping and visualization) and Harzing's Publish or Perish (performance indicators)

Table I positions the present work relative to recent bibliometric studies by comparing their coverage periods, database selections, keyword strategies, analytical focus and toolchains. Prior studies [10], [11] and [13] vary in scope and objectives, including broad misinformation-dissemination mapping and incremental-learning-oriented surveys and they often rely on Web of Science or a combination of databases, with different preprocessing pipelines. In contrast, the current study targets a narrower technical stream by focusing specifically on Transformer-based architectures for misinformation detection. It adopts a consistent Scopus-only retrieval strategy covering 2017–2026 (as of 3 March 2026). Scopus was selected as the sole data source because it offers broad coverage of computer science and engineering venues and provides structured bibliographic metadata required for performance analysis and science mapping (e.g., author affiliations, source titles, references and citation counts) under a consistent indexing policy. Using a single database also reduces duplication and record-matching errors that can arise when merging records from multiple sources with different metadata schemas.

Nevertheless, Scopus-only retrieval may omit publications indexed exclusively elsewhere and may underrepresent some regional outlets. Accordingly, the results should be interpreted as representing the Scopus-indexed landscape during the retrieval period. Future work may triangulate Scopus with Web

of Science or other databases to test the stability of the identified trends and networks.

The keyword formulation also differs by explicitly combining misinformation terms with Transformer-related terms, enabling the dataset to capture Transformer-centric model families and related evaluation settings rather than general diffusion or platform-level discourse. Methodologically, the study integrates performance analysis with science mapping, supported by the tool configuration reported in Table I and the retrieval workflow shown in Fig. 1.

A. Data Source and Search Scope

Records were retrieved from Scopus on 3rd March 2026 using the TITLE-ABS-KEY fields. The search covered all years, was restricted to English, and applied no restrictions on source type or document type. The workflow is summarized in Fig. 1. The query is intentionally inclusive, so the retrieved set may contain adjacent high-impact LLM-related work that shapes misinformation threat models even when it is not a detection-model paper.

B. Search Strategy and Query Design

The query combined (1) misinformation-related terms (e.g., misinformation, disinformation, fake news, rumor) with (2) Transformer-related terms (e.g., Transformer, attention mechanism, BERT, RoBERTa, LLM). The complete Boolean query is provided in Fig. 1.

C. Study Selection and Dataset Construction

The search returned 3,637 records. Screening was limited to database-level filters (TITLE-ABS-KEY and English). No post-retrieval exclusions were applied; therefore, 0 records were removed, and 3,637 documents were included (Fig. 1). To clarify the dataset scope, the 3,637 records are treated as a query-defined Scopus corpus rather than a manually screened systematic review corpus. The query was designed to capture both detection-oriented studies and related Transformer/LLM work that informs misinformation threat models and research trends. Therefore, no post-retrieval topical exclusion was applied, and the dataset should be interpreted as the Scopus-indexed landscape captured by the stated search strategy, not as a set limited only to detection-architecture papers.

D. Data Export and Preprocessing

All retrieved records were exported from Scopus, including available bibliographic metadata such as publication year, document title, authors, affiliations, source title, abstracts, author keywords, indexed keywords, cited references, and citation counts. Before analysis, the dataset was checked for basic consistency to reduce noise in network construction. This included harmonizing common variants of author names and keywords, where applicable, and consolidating obvious duplicates (for example, singular and plural forms or abbreviations representing the same concept). The cleaned files were then prepared in formats compatible with the selected tools.

E. Bibliometric Tools and Analysis Procedures

VOSviewer (v1.6.20) was used to construct and visualize co-authorship networks involving authors, institutions, and countries, as well as keyword co-occurrence and overlay maps. Citation indicators were examined using Harzing's Publish or Perish to support publication-level and source-level impact analysis.

F. Outcome Measures and Reporting

Results are reported using (1) performance indicators describing publication growth, major contributors, leading sources, and highly cited documents, and (2) science mapping outputs that visualize collaboration structure and thematic organization. The overall workflow is summarized in Fig. 1.

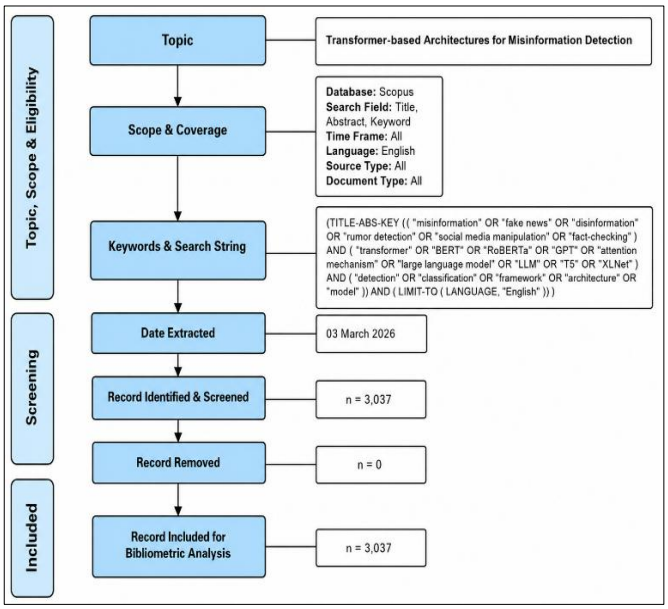


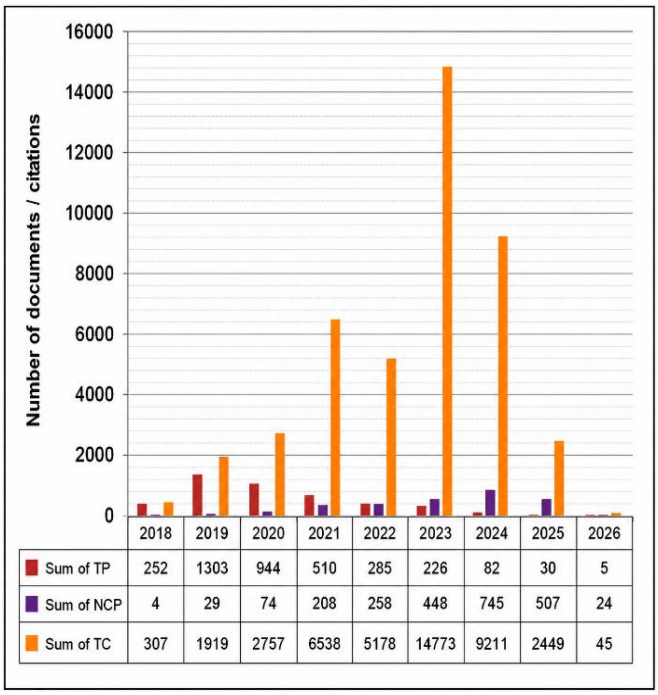
Fig. 1. Flowchart illustrating the literature search and data analysis process in the study.

III. Results and Discussion

A. Publication Output and Citation Trend

The temporal pattern of research activity and impact is described using three indicators: total publications (TP), number of cited publications (NCP), and total citations (TC). As shown in Fig. 2, publication output increases from 252 papers in 2018 to a peak of 1,303 papers in 2019, followed by 944 papers in 2020. Output then declines in subsequent years (510 in 2021, 285 in 2022, 226 in 2023, 82 in 2024, 30 in 2025 and 5 in 2026). The reduction in the most recent years is consistent with a shorter publication and indexing window, as the dataset was extracted in early 2026.

Indexing delay can explain part of the reduction in 2024–2026, but it does not fully account for the sustained decline between 2019 and 2023. A plausible interpretation is that the 2019 peak reflects an early surge of BERT-family adoption in misinformation detection, while subsequent work increasingly shifted toward adjacent framings (e.g., fact-checking, claim verification, stance and content authenticity) and reporting conventions that do not always foreground the same Transformer keywords in TITLE-ABS-KEY fields. Accordingly, the post-2019 decline should be interpreted as a query- and indexing-conditioned pattern within the retrieved Scopus dataset rather than a definitive indicator of reduced Transformer-based research activity.



	2018	2019	2020	2021	2022	2023	2024	2025	2026
Sum of TP	252	1303	944	510	285	226	82	30	5
Sum of NCP	4	29	74	208	258	448	745	507	24
Sum of TC	307	1919	2757	6538	5178	14773	9211	2449	45

Note: TP = Total Publications; NCP = Cited Publications; TC = Total Citations.

Fig. 2. Annual publication output and citation impact.

Citation indicators exhibit a lag relative to publication volume. TC rises from 307 (2018) and 1,919 (2019) to 2,757 (2020) and 6,538 (2021), remains high in 2022 (5,178) and peaks in 2023 (14,773). TC then decreases in 2024 (9,211), drops further in 2025 (2,449) and declines in 2026 (45), reflecting limited time for recent outputs to accumulate citations. A similar pattern is observed for NCP, which increases from 4 (2018) and 29 (2019) to 74 (2020) and 208 (2021), reaches 258 (2022) and 448 (2023), peaks in 2024 (745) and declines in 2025 (507) and 2026 (24).

Overall, the year-wise patterns indicate that publication volume is concentrated in earlier years within this dataset, while citation impact becomes more visible after a maturation period. This lag implies that citation-based influence should be interpreted using multi-year windows rather than recent-year counts alone, since newer publications have not had sufficient time to accumulate citations. In practical terms, the most recent declines are better viewed as recency and indexing effects than as evidence of reduced scholarly attention. This answers RQ1 by describing how publication output and citation impact evolve. Having established this temporal profile, the next subsection evaluates the impact more directly by identifying influential publications and sources.

B. Citation Impact of Publications and Sources

Citation impact is assessed at two levels: publication-level influence, using citation indicators and highly cited papers and source-level influence, indicating where influential work is disseminated and where recent attention is concentrated. Publication-level impact is summarized in Table II.

As shown in Table II, the dataset contains 3,637 papers and 43,207 citations, with an average of 11.88 citations per paper, an h-index of 82, and a g-index of 154. These indicators suggest broad influence with a highly cited core, which motivates a closer look at the publications driving this impact. The most cited publications are listed in Table III. Because the dataset follows a query-defined bibliometric scope, Table III is interpreted as showing the most-cited publications within the retrieved corpus rather than as a list restricted to detection-architecture papers. Some records are not detection papers in a narrow methodological sense, but they are retained because they represent influential Transformer- and LLM-related work that shapes current misinformation research, especially around generative AI, information integrity and emerging threat models.

TABLE II. CITATION METRICS

Metrics	Data
Papers	3637
Number of Citations	43207
Years	8
Citations per Year	5400.88
Citations per Paper	11.88
Cites_Author	11525.38
Papers_Author	1170.09
Authors_Paper	4.1
h_index	82
g_index	154

Table III should be read as reflecting the most influential publications within the retrieved scope, rather than as a list restricted to detection-architecture papers. It indicates two major citation drivers. Several highly cited items focus on LLMs and ChatGPT, reflecting strong cross-domain attention to generative AI and its implications for information integrity and high-stakes settings. A second stream comprises methodological contributions, including detection baselines and multimodal frameworks that are repeatedly cited as reference implementations or comparative anchors in subsequent studies. This split suggests that citation impact is shaped by both the rapid uptake of LLM-related discourse and the persistent reliance on foundational detection frameworks. Importantly, this pattern implies that influence in the domain is determined not only by advances in modelling but also by how research engages with emerging threat models and deployment-relevant concerns. As a result, studies that connect detection performance with robustness, transparency, or real-world risk contexts are more likely to achieve sustained citation visibility.

Source-level dissemination is examined in Fig. 3, which contrasts Total Publications (TP) with citations in 2025 across leading venues grouped by source type. While Table II and Table III provide a publication-level view of impact, it is also important to examine the sources/venues through which this research is disseminated.

TABLE III. TOP 10 CITED PUBLICATIONS

No.	Authors	Title	Cites	Cites per Year
1	Y.K. Dwivedi et al. (2023) [15]	So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy	3217	15.67
2	J. Clusmann et al. (2023) [16]	The future landscape of large language models in medicine	695	19
3	R.K. Kaliyar et al. (2021) [17]	FakeBERT: Fake news detection in social media with a BERT-based deep learning approach	673	11
4	S. Singhal et al. (2019) [18]	SpotFake: A multi-modal framework for fake news detection	553	10
5	L. De Angelis et al. (2023) [19]	ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health	501	12.17
6	A. Abd-alrazaq et al. (2023) [20]	Large Language Models in Medical Education: Opportunities, Challenges and Future Directions	438	7.44
7	J. Dempere et al. (2023) [21]	The impact of ChatGPT on higher education	343	9.83
8	S. Harrer (2023) [22]	Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine	334	13.5
9	A. Choudhury, H. Shamszare (2023) [23]	Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis	326	5.88
10	P. Qi, J. Cao, T. Yang, J. Guo, J. Li (2019) [24]	Exploiting multi-domain visual information for fake news detection	324	4.6

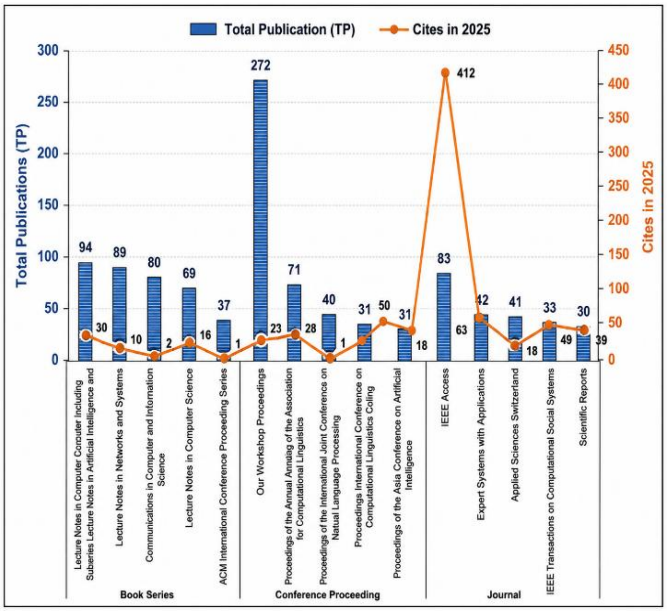


Fig. 3. Publication output (TP) and citations in 2025 across leading sources.

In Fig. 3, the highest-output venues differ by source type. Within Book Series, Lecture Notes in Computer Science (including LNAI/LNBI) leads output (TP = 94), followed by Lecture Notes in Networks and Systems (TP = 89), and Communications in Computer and Information Science (TP = 80). Within Conference Proceedings, CEUR Workshop Proceedings dominates (TP = 272), followed by the ACL Annual Meeting Proceedings (TP = 71). Within Journals, IEEE Access records the highest output (TP = 83), followed by Expert Systems with Applications (TP = 42), and Applied Sciences (Switzerland) (TP = 41). Citations in 2025 are used as a fixed one-year window to reduce the advantage of older publications and to highlight venues receiving current attention within the same citation period.

Taken together, the source-level profile separates dissemination capacity (TP) from recent visibility (2025 citations). This distinction is useful because high output does not necessarily imply high short-window citation uptake and it provides a venue-focused complement to publication-level impact indicators. Table II, Table III and Fig. 3, therefore, provide publication and venue-level evidence of impact. This answers RQ2 by identifying the most influential publications and the main sources through which impactful work is disseminated. A limitation is that citations in a single-year window and source-level counts are sensitive to indexing coverage and retrieval scope. Accordingly, these indicators should be interpreted as reflecting venue visibility within the Scopus-indexed records captured by the query and retrieval date.

C. Subject Area Distribution and Interdisciplinarity

The distribution of publications across Scopus subject areas is used to characterize the disciplinary base of Transformer-based misinformation detection research and to assess the degree of interdisciplinarity. Because Scopus can assign a single document to multiple subject areas, percentages may sum to more than 100 per cent; the overlap reflects multi-disciplinary indexing rather than duplication in the dataset.

As shown in Table IV, Computer Science accounts for 3,091 publications (84.99%), indicating that the field is primarily anchored in computing and NLP research. Engineering accounts for 1,018 publications (27.99%), followed by Mathematics with 796 (21.89%), aligning with the emphasis on model development, optimization and evaluation methodology. Beyond the technical core, Social Sciences records 560 publications (15.40%), Decision Sciences contributes 404 (11.11%) and Medicine accounts for 324 (8.91%), reflecting sustained interest in misinformation as a socio-technical problem with decision and health implications.

The overall pattern is summarized visually in Fig. 4, which focuses on the top 10 subject areas and clearly shows the dominance of Computer Science while also highlighting meaningful participation from adjacent disciplines. Taken together, these results suggest that Transformer-based misinformation detection is rooted in computer science research but increasingly shaped by interdisciplinary applications, particularly in social and health domains.

TABLE IV. SUBJECT AREA

Subject Area	TP	%
Computer Science	3,091	84.99%
Engineering	1,018	27.99%
Mathematics	796	21.89%
Social Sciences	560	15.40%
Decision Sciences	404	11.11%
Medicine	324	8.91%
Arts and Humanities	206	5.66%
Materials Science	185	5.09%
Physics and Astronomy	140	3.85%
Business, Management and Accounting	106	2.91%
Multidisciplinary	77	2.12%
Energy	72	1.98%
Chemical Engineering	47	1.29%
Neuroscience	44	1.21%
Psychology	39	1.07%
Health Professions	38	1.04%
Biochemistry, Genetics and Molecular Biology	36	0.99%
Environmental Science	34	0.93%
Chemistry	18	0.49%
Agricultural and Biological Sciences	16	0.44%
Economics, Econometrics and Finance	15	0.41%
Dentistry	14	0.38%
Nursing	13	0.36%
Earth and Planetary Sciences	10	0.27%
Pharmacology, Toxicology and Pharmaceutics	7	0.19%
Immunology and Microbiology	3	0.08%
Veterinary	1	0.03%

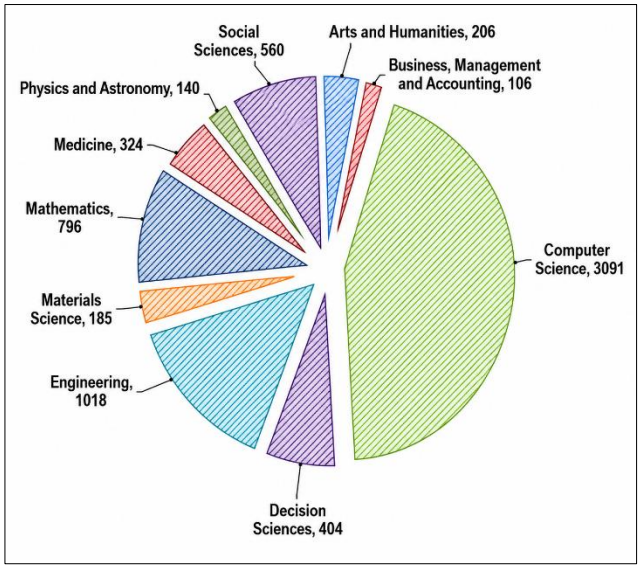


Fig. 4. Top ten subject areas.

Overall, the subject-area profile suggests that Transformer-based misinformation detection is anchored in Computer Science and draws substantial contributions from Engineering and Mathematics, reflecting a technical emphasis on model design, optimization and learning foundations. At the same time, the visible presence of Social Sciences, Decision Sciences and Medicine indicates that the field is increasingly shaped by applied and societal contexts, where misinformation is studied not only as a classification problem but also as a phenomenon involving human behavior, risk and domain-specific consequences. This interdisciplinary pattern provides useful context for interpreting the thematic clusters and collaboration networks in later analyses by explaining why research topics often connect algorithmic advances with application-driven concerns. These findings directly address RQ6 by showing that the field remains anchored in technical disciplines while drawing increasing attention from applied and societal domains.

Having established the disciplinary footprint of the literature, the next subsection examines country-level publication output to characterize the geographic distribution of research activity.

D. Country Contribution and Geographic Distribution

Country-level publication output is used to identify the geographic concentration of research activity in Transformer-based misinformation detection. The top contributing countries are summarized in Table V and their spatial distribution is illustrated in Fig. 5.

TABLE V. TOP 10 COUNTRIES WITH THE MOST TOTAL PUBLICATIONS

Country	TP	%
China	839	23.07%
United States	708	19.47%
India	645	17.73%
United Kingdom	199	5.47%
Germany	169	4.65%
Italy	142	3.90%
Australia	126	3.46%
Spain	116	3.19%
Saudi Arabia	115	3.16%
Canada	113	3.11%

As shown in Table V, publication output is concentrated among a small set of countries. China ranks first with 839 publications (23.07%), followed by the United States with 708 (19.47%) and India with 645 (17.73%). The next contributors are the United Kingdom (199, 5.47%), and Germany (169, 4.65%), followed by Italy (142, 3.90%), and Australia (126, 3.46%). Spain (116, 3.19%), Saudi Arabia (115, 3.16%), and Canada (113, 3.11%) complete the top ten.

Fig. 5 shows the same ranking, where shading intensity reflects publication volume. The concentration in a small number of countries suggests that a limited set of national research ecosystems drives output. At the same time, the presence of contributors across Asia, North America, Europe and the Middle East indicates broad international participation.

This distribution provides context for later analysis of collaboration, since geographic concentration can influence co-authorship density and cross-border research ties. This answers RQ3 by identifying the countries that contribute most to the literature.

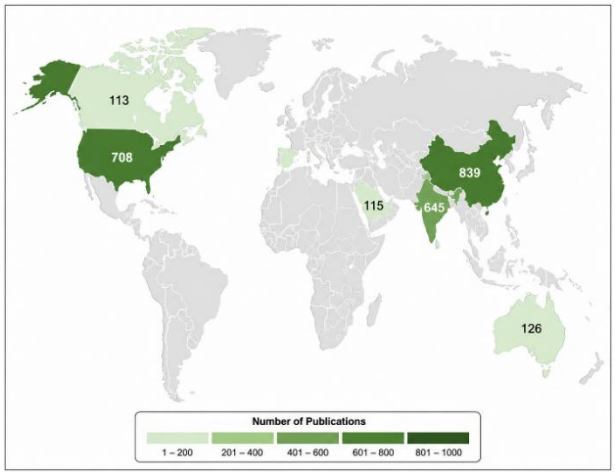


Fig. 5. Geographic distribution of the top 10 countries by total publications.

E. Institutional Contribution

Institutional-level publication output is summarized to identify the primary organizational hubs contributing to Transformer-based misinformation detection research. The distribution of the most productive institutions is shown in Fig. 6.

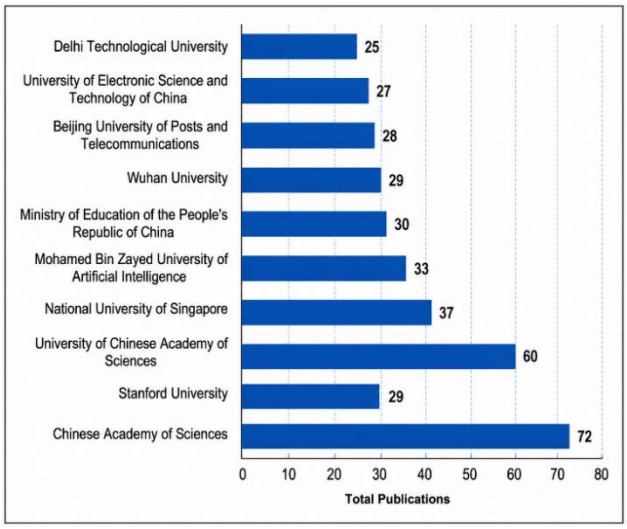


Fig. 6. Top 10 institutions by total publications.

As shown in Fig. 6, output is concentrated in a small set of institutions. The Chinese Academy of Sciences (CAS) ranks first with 72 publications, followed by the University of Chinese Academy of Sciences (UCAS) with 60 publications. These two entities are closely linked within the same research ecosystem. Still, they are reported separately because they are indexed as distinct affiliations and authors frequently list research institute and university affiliations independently. In the next tier, the National University of Singapore records 37 publications and Mohamed Bin Zayed University of Artificial Intelligence records 33 publications. Additional contributors include the Ministry of Education of the People's Republic of China (30), Wuhan University (29), Stanford University (29), Beijing University of Posts and Telecommunications (28), University of Electronic Science and Technology of China (27), and Delhi Technological University (25).

Overall, the institutional profile indicates that publication activity is strongly represented by China-based organizations, with major contributors from the United States, Singapore, UAE, and India. This answers RQ3 by identifying the institutions with the highest publication output in the dataset. Having established institutional concentration, the next subsection examines author productivity and collaboration patterns to characterize how research communities are organized at the researcher level.

F. Author Productivity and Collaboration Patterns

The collaboration structure is illustrated by the co-authorship network in Fig. 7, complemented by the productivity ranking of leading authors in Table VI. The most productive authors are summarized in Table VI and the co-authorship network is illustrated in Fig. 7.

TABLE VI. LEADING AUTHORS BY TOTAL PUBLICATIONS

Author full names	Affiliations	Author	TP
Nakov, Preslav Ivanov	Mohamed Bin Zayed University of Artificial Intelligence	United Arab Emirates	33
Alam, Firoj	Hamad Bin Khalifa University	Qatar	17
Zhang, Min	Peking University	China	12
Martín, Antonio	Universidad Politécnica de Madrid	Spain	11
Camacho, David	Universidad Autónoma de Madrid	Spain	11
Hoque, Md. Moshiur	University of Dhaka	Bangladesh	11
Ma, Jing	The Hong Kong Polytechnic University	Hong Kong	10
Madasamy, Anand	Indian Institute of Technology	India	10
Wasim, Muhammad	Lahore University of Management Sciences	Pakistan	10
Alghamdi, Jaber	King Abdulaziz University	Saudi Arabia	10

Note: TP=total number of publications

As shown in Table VI, publication output is concentrated among a small group of authors. Nakov, Preslav Ivanov, ranks first with 33 publications, followed by Alam, Firoj, with 17 and Zhang, Min, with 12. The next contributors include Martín, Antonio (11), Camacho, David (11) and Hoque, Md. Moshiur (11). Several authors contribute at a comparable level with 10 publications each, including Ma, Jing, Madasamy, Anand, Wasim, Muhammad and Alghamdi, Jaber. This distribution indicates that author productivity is top-heavy, with a small set of frequent contributors and a broader group of steady contributors.

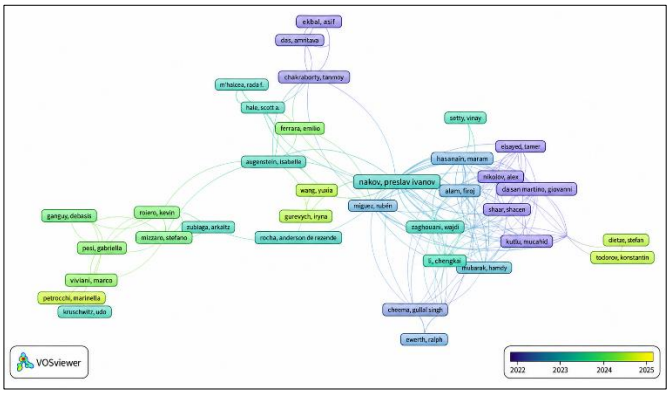


Fig. 7. Co-authorship network visualization.

The co-authorship network in Fig. 7 provides a structural view of collaboration. Nodes represent authors and links indicate co-authored publications. Larger nodes indicate greater publication presence in the dataset, while denser link patterns indicate repeated collaboration. The network forms distinct clusters, indicating collaboration communities in which authors frequently publish together. Some authors are embedded within tightly connected clusters, consistent with stable research groups. In contrast, others occupy bridging positions between clusters, suggesting cross-group collaboration pathways that can facilitate the spread of methods and datasets across the field. The color gradient adds a temporal signal by distinguishing earlier versus more recent activity within the displayed period.

Taken together, the author productivity ranking and co-authorship structure provide evidence of concentrated productivity and cluster-based collaboration patterns. This answers RQ4 by identifying the most productive authors and the dominant collaboration structure in the field. Having established the leading contributors and collaboration patterns, the next subsection maps the literature's thematic structure using keyword co-occurrence and overlay analysis.

G. Keyword Co-occurrence and Thematic Evolution

The thematic structure of Transformer-based misinformation detection research is characterized through keyword co-occurrence patterns and their temporal evolution. The co-occurrence structure is shown in Fig. 8, while the overlay visualization in Fig. 9 highlights topic recency based on average publication year. Keyword labels such as *fake news* and *disinformation* are reported as indexed terms in the dataset and are interpreted within the broader scope of misinformation detection defined in the Introduction.

As shown in Fig. 8, the literature is organized around a dense core of high-frequency keywords, with strong connectivity among Natural Language Processing, social media and misinformation-related terms. This configuration indicates that a substantial portion of the field is framed as NLP-based detection in social media environments. The map also shows connected sub-themes that reflect methodological diversification, including multimodal and cross-modal directions (e.g., terms associated with multimodal fusion and vision-based processing) and application-driven contexts such as COVID-19, which continues to appear as an anchoring topic. The presence of both baseline classifier terms and Transformer-family model terms indicates that comparative evaluation against conventional baselines remains common alongside modern representation learning approaches.

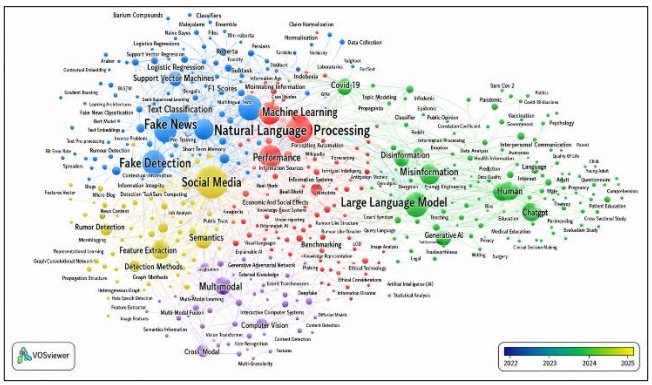


Fig. 8. Keyword co-occurrence map.

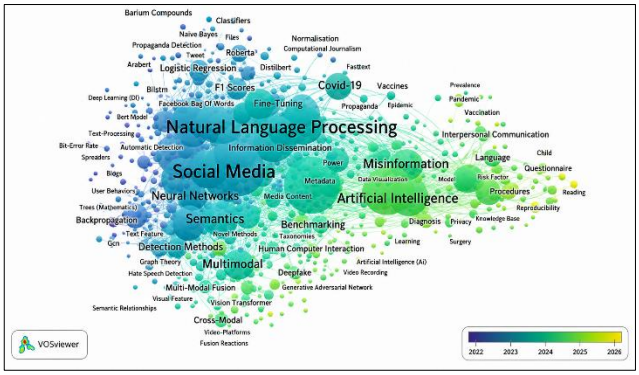


Fig. 9. Keyword overlay visualization map.

The overlay view in Fig. 9 provides a time-sensitive interpretation of keyword activity. Earlier topics are more strongly associated with general detection and evaluation vocabulary, while more recent activity shifts toward keywords linked to contemporary AI developments and deployment-oriented concerns. The appearance of terms related to generative systems and research-practice issues (e.g., reproducibility and procedure-oriented keywords) suggests that recent work is expanding beyond benchmark performance toward questions of reliability, reporting rigor and applicability in real-world conditions. At the same time, core detection topics remain active, indicating continuity in the main problem setting while also highlighting emerging directions. This temporal shift implies that evaluation designs may need to move beyond static benchmark settings toward protocols that reflect emerging threat models, including synthetic content and multimodal manipulation. It also indicates that research visibility is increasingly linked to deployment-oriented themes, which can guide future work toward robustness, transparency and real-world validation.

Specifically, the earlier activity in Fig. 9 is more closely associated with classical detection and evaluation vocabulary, such as fake news, text classification, support vector machines, logistic regression, rumor detection and F1 score. In contrast, more recent keywords include ChatGPT, generative AI, prompt engineering, reproducibility, Vision Transformer (ViT) and

multimodal fusion, indicating a shift toward LLM-driven authenticity concerns, multimodal settings and deployment-oriented research practices.

Overall, the combined evidence from Fig. 8 to Fig. 9 identifies dominant thematic clusters and clarifies how attention has shifted over time. This answers RQ5 by characterizing the domain's thematic structure and its evolution through keyword co-occurrence and overlay analysis. Having established the thematic structure and its temporal shifts, the next subsection synthesizes research gaps and future directions derived from the observed trends in methods, applications and evaluation practices.

A limitation of keyword co-occurrence mapping is its dependence on author keywording practices and indexing variation, which can affect term visibility and apparent connectivity. Therefore, Fig. 8 should be read as indicating dominant thematic signals in the retrieved dataset rather than an exhaustive representation of all topics.

H. Research Gaps and Future Directions

The thematic structure and temporal overlay indicate that the field is expanding rapidly, yet several gaps remain that limit the reliability and deployability of Transformer-based misinformation detection. The key gaps and recommended future directions derived from the bibliometric findings are summarized in Table VII.

TABLE VII. RESEARCH GAPS AND FUTURE DIRECTIONS IN TRANSFORMER-BASED MISINFORMATION DETECTION

Gap area	Evidence from this study	Why it matters	Future directions (recommended)
Evaluation of realism and robustness	Keyword maps emphasize benchmark and evaluation terms (Fig. 8) and recent attention to reproducibility (Fig. 9).	Benchmark gains may not translate to deployment when data shifts across time, topics and platforms.	Use cross-domain and cross-platform evaluation, temporal holdout testing, robustness checks under distribution shift and transparent reporting of splits and baselines.
LLM-generated misinformation	Highly cited works are dominated by publications on LLMs/ChatGPT (Table III).	Detection pipelines often assume human-authored content, while LLM text can be prompt-sensitive and rapidly evolving.	Develop detection strategies for synthetic text (prompt variation, paraphrasing, mixed human–AI authorship), integrate verification signals and evaluate against evolving generator models.
Multimodal misinformation detection	Multimodal themes appear as a connected topic area (Fig. 8) and several influential publications are multimodal (Table III).	Real-world misinformation often combines text and images, so text-only models may miss important signals or be susceptible to context manipulation.	Expand multimodal datasets, evaluate cross-modal consistency and robustness and design fusion architectures that handle missing or manipulated modalities.
Explainable and trustworthy detection	Explainability-related directions appear as emerging concerns alongside deployment-oriented themes (Fig. 9).	High-stakes settings require interpretability and trust, not only accuracy, to support verification and decision-making.	Incorporate explanation methods (faithful rationales, uncertainty estimates), audit bias and failure modes and evaluate explanation quality with human-centered protocols.
Cross-lingual and low-resource settings	The thematic core is strongly anchored in mainstream NLP and social media detection terminology (Fig. 8), suggesting a dominance of high-resource settings.	Models trained on English-heavy benchmarks often generalize poorly to other languages, dialects and code-switching contexts.	Build multilingual and low-resource benchmarks, develop adaptation methods (transfer, multilingual pretraining, PEFT) and evaluate fairness across languages and regions.
Reproducibility and transparency	"Reproducibility" emerges as a recent keyword signal in the overlay map (Fig. 9).	Without consistent reporting and reusable pipelines, performance improvements may reflect experimental variance rather than real advances.	Release code and data splits, standardize preprocessing and baselines, report hyperparameters and adopt reproducible evaluation toolchains.

Taken together, Table VII indicates that recent progress is not limited by model capacity alone, but by evaluation realism, modality coverage and reporting discipline. This implies that future contributions will be more defensible when they adopt shift-aware evaluation protocols and provide reproducible pipelines alongside improvements in accuracy.

First, the literature remains strongly benchmark-driven, with frequent emphasis on standard classification performance and metric reporting, as reflected by the prominence of evaluation terms in the keyword map and its overlay evolution (Fig. 8 and Fig. 9). While benchmark progress supports comparability, the field would benefit from evaluation protocols that better reflect real-world distribution shifts over time, across platforms and across topics. Future work should prioritize cross-domain and cross-platform testing, temporal generalization and robustness assessment in the face of changing narratives and emerging events. Prior evidence also shows that dataset limitations, feature representation choices and simplistic fusion strategies remain recurring failure points in misinformation detection systems, reinforcing the need for evaluation designs that reflect these constraints [25], [26].

Second, the rise of LLMs and ChatGPT is reshaping the misinformation landscape and requires explicit methodological responses. Highly cited publications emphasize the influence and societal implications of generative conversational AI [11], including risks in sensitive domains such as public health and education [15]-[18]. However, detection pipelines often assume human-authored content and may not be calibrated to synthetic text produced by variations in prompts and model updates. Future studies should develop detection and verification strategies specifically designed for LLM-generated misinformation, including robustness under prompt manipulation, synthetic paraphrasing and mixed human–AI authorship settings, while also considering human trust and adoption behaviors that shape how such content is consumed [19].

Third, multimodal misinformation remains under-addressed relative to its prevalence in online ecosystems. Although

multimodal cues have emerged as a theme and influential works have demonstrated the value of integrating visual information [14], [20], the literature still concentrates heavily on text-only inputs. Future research should move toward multimodal and cross-modal architectures evaluated on realistic mixed-media datasets, including scenarios with incomplete, noisy, or manipulated visual context. In parallel, explainable and trustworthy AI should be treated as a first-class requirement, particularly for high-stakes environments and public-facing deployments. Recent work on explainable multilingual and multimodal detection supports this direction [6], [7], but broader adoption remains limited. Beyond modality fusion, incorporating affective signals from news content and user responses has also been shown to improve detection performance on social platforms, suggesting an additional direction for feature enrichment in real-world settings [27].

Fourth, low-resource and cross-lingual settings remain comparatively underrepresented. The keyword landscape suggests dominance of mainstream NLP terms and widely used benchmarks, which often privilege English and high-resource contexts. To improve real-world applicability, future work should invest in multilingual datasets, domain adaptation techniques, and evaluation settings that emphasize fairness and generalization across languages, including code-switching and regional dialects [28]. Noisy, non-standard social media language and spelling variation can further degrade cross-lingual transfer, motivating normalization-aware pipelines and curated resources for low-resource settings [29], [30].

Finally, reproducibility and methodological transparency require stronger emphasis. The overlay map highlights "reproducibility" as a more recent concern (Fig. 9), indicating growing awareness of the need for reporting rigor. Future studies should adopt clearer documentation of preprocessing, data splits, hyperparameter configurations, and baseline selection. They should publish reusable evaluation pipelines to ensure that progress reflects genuine methodological improvement rather than experimental variance [31].

In summary, these gaps collectively answer RQ7 by identifying priorities for advancing reliable and deployable misinformation detection: 1) realistic evaluation and robustness, 2) explicit handling of LLM-generated misinformation, 3) stronger multimodal and explainable frameworks, 4) broader cross-lingual coverage, and 5) improved reproducibility and transparency.

IV. Conclusion

This study presents a Scopus-based bibliometric analysis of Transformer-based misinformation detection, drawing on 3,637 records, and provides an evidence-driven account of how the field is currently organized and where it is moving. The results show that publication output is concentrated in conference proceedings and journals and that the research base is firmly anchored in Computer Science, with sustained contributions from applied disciplines that reflect deployment-facing concerns. At the contributor level, productivity is concentrated among a small set of leading countries, institutions and co-authorship communities, indicating that research capacity is not evenly distributed. Citation metrics further confirm broad influence with a distinct, highly cited core. At the same time, keyword mapping identifies a stable focus on NLP-driven misinformation detection in social media and a clear recent shift toward multimodal pipelines and LLM-related themes.

The primary contribution of this study is a consolidated, reproducible map of productivity, influence and thematic evolution that reduces fragmentation in the literature and supports a more defensible positioning for future work. Future research may further deepen intellectual-structure analysis through citation, co-citation, and bibliographic coupling networks, refine time-sliced thematic tracking to capture topic turnover, and strengthen external validity through reproducible evaluation protocols, cross-domain and cross-platform robustness testing, and expanded coverage of low-resource and cross-lingual settings.

Acknowledgment

This study is a research product supported and funded by the Fundamental Research Grant Scheme (FRGS) of the Ministry of Higher Education Malaysia (Code: FRGS/1/2022/ICT02/ UKM/02/3).

Declaration of Generative AI in Scientific Writing

Generative AI tools were used only for language editing to improve clarity and readability. All AI-assisted revisions were reviewed and applied by the authors. The authors take full responsibility for the accuracy and integrity of the work. No AI tools were used for conceptualization, study design, data analysis, result interpretation or the development of the scientific content. In accordance with journal and ethical guidelines, no AI tool is listed as an author or co-author.

References

[1] G. Nogara et al., "A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch," Online Social Networks and Media, vol. 51, Art. no. 100342, 2026, doi: 10.1016/j.osnem.2025.100342.

[2] C. Bono, N. Liu, G. Russo and F. Pierri, "Self-moderation in the decentralized era: decoding blocking behavior on Bluesky," arXiv:2505.01174, 2025. [Online]. Available: https://arxiv.org/abs/2505.01174

[3] D. Mouratidis, A. Kanavos and K. Kermanidis, "From misinformation to insight: Machine learning strategies for fake news detection," Information, vol. 16, no. 3, Art. no. 189, 2025, doi: 10.3390/info16030189.

[4] Q. Su, M. Wan, X. Liu and C.-R. Huang, "Motivations, methods and metrics of misinformation detection: An NLP perspective," Natural Language Processing Research, vol. 1, nos. 1-2, Art. no. 1, 2020, doi: 10.2991/nlpr.d.200522.001.

[5] A. O. Mohamed, I. Asim, I. Eltayeb, A. Ruslan and N. E. Jeffry, "Fake news detection: A review of conventional and state-of-the-art approaches," Journal of Applied Science, Technology and Computing, vol. 2, no. 1, pp. 43-54, 2025. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jastec/article/view/23956

[6] C. Wang, M. Gao, Z. Wang, J. Yin, K. Shu and C. Lin, "Prompt-induced linguistic fingerprints for LLM-generated fake news detection," arXiv:2508.12632, 2025. [Online]. Available: https://arxiv.org/abs/2508.12632

[7] L. Pangtey, A. Bhatnagar, S. Bansal, S. S. Dar and N. Kumar, "Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions," arXiv:2505.08464, 2025. [Online]. Available: https://arxiv.org/abs/2505.08464

[8] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash and S. Jadhav, "Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation," Frontiers in

Artificial Intelligence, vol. 8, Art. no. 1690616, 2025, doi: 10.3389/frai.2025.1690616.

[9] C. A. Anser Pasha and S. Sampath, "X-FakeNet: Explainable multimodal fusion with meta-reasoning for fake news detection," International Journal of Intelligent Engineering and Systems, vol. 18, no. 11, pp. 645-659, 2025, doi: 10.22266/ijies2025.1231.40.

[10] G. Xu, M. Qian and L. Meng, "Misinformation dissemination on social media: Key research themes and evolutionary paths between 2013 and 2023," Humanities and Social Sciences Communications, vol. 12, Art. no. 1775, 2025, doi: 10.1057/s41599-025-06067-1.

[11] S. Wang, F. Su, L. Ye and Y. Jing, "Disinformation: A bibliometric review," International Journal of Environmental Research and Public Health, vol. 19, no. 24, Art. no. 16849, 2022, doi: 10.3390/ijerph192416849.

[12] A. Alarabid, M. Ali, E. B. Abbad, A. Al Bloushi and H. Heba, "Leveraging large language models for misinformation detection: A focus on public health misinformation on social media," in Lecture Notes in Networks and Systems, vol. 1539, pp. 561-573, 2026, doi: 10.1007/978-981-96-9248-4_42.

[13] Y. N. Barve and P. Mulay, "Bibliometric survey on incremental learning in text classification algorithms for false information detection," Library Philosophy and Practice (e-journal), Art. no. 4560, 2020. [Online]. Available: https://digitalcommons.unl.edu/libphilprac/4560/

[14] S. Rastogi and D. Bansal, "A review on fake news detection 3T's: Typology, time of detection, taxonomies," International Journal of Information Security, vol. 22, no. 1, pp. 177-212, 2023, doi: 10.1007/s10207-022-00625-3.

[15] Y. K. Dwivedi et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," International Journal of Information Management, vol. 71, Art. no. 102642, 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[16] J. Clusmann et al., "The future landscape of large language models in medicine," Communications Medicine, vol. 3, no. 1, 2023, doi: 10.1038/s43856-023-00370-1.

[17] R. K. Kaliyar, A. Goswami and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools and Applications, vol. 80, no. 8, pp. 11765-11788, 2021, doi: 10.1007/s11042-020-10183-2.

[18] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in Proc. IEEE Int. Conf. Multimedia Big Data (BigMM), 2019, pp. 39-47, doi: 10.1109/BigMM.2019.00-44.

[19] L. De Angelis et al., "ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health," Frontiers in Public Health, vol. 11, Art. no. 1166120, 2023, doi: 10.3389/fpubh.2023.1166120.

[20] A. Abd-alrazaq et al., "Large language models in medical education: Opportunities, challenges and future directions," JMIR Medical Education, vol. 9, Art. no. e48291, 2023, doi: 10.2196/48291.

[21] J. Dempere, K. Modugu, A. Allam and L. K. Ramasamy, "The impact of ChatGPT on higher education," Frontiers in Education, vol. 8, Art. no. 1206936, 2023, doi: 10.3389/feduc.2023.1206936.

[22] S. Harrer, "Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine," eBioMedicine, vol. 90, Art. no. 104512, 2023, doi: 10.1016/j.ebiom.2023.104512.

[23] A. Choudhury and H. Shamszare, "Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis," Journal of Medical Internet Research, vol. 25, Art. no. e47184, 2023, doi: 10.2196/47184.

[24] P. Qi, J. Cao, T. Yang, J. Guo and J. Li, "Exploiting multi-domain visual information for fake news detection," in Proc. IEEE Int. Conf. Data Mining (ICDM), 2019, pp. 518-527, doi: 10.1109/ICDM.2019.00062.

[25] S. Kh. Hamed, M. J. Ab Aziz and M. R. Yaakub, "A review of fake news detection models: Highlighting the factors affecting model performance and the prominent techniques used," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 7, pp. 379-390, 2023.

[26] S. Kh. Hamed, M. J. Ab Aziz and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation and data fusion," Heliyon, vol. 9, Art. no. e20382, 2023, doi: 10.1016/j.heliyon.2023.e20382.

[27] S. Kh. Hamed, M. J. Ab Aziz and M. R. Yaakub, "Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments," Sensors, vol. 23, Art. no. 1748, 2023, doi: 10.3390/s23041748.

[28] X. Wang, W. Zhang and S. Rajtmajer, "Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey," arXiv:2410.18390, 2024. [Online]. Available: https://arxiv.org/abs/2410.18390

[29] S. N. A. N. Ariffin and S. Tiun, "Rule-based text normalization for Malay social media texts," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 10, pp. 156-162, 2020.

[30] Z. Mansur, N. Omar, S. Tiun and E. M. Alshari, "A hybrid BERT-based normalization framework for hate speech detection," Asia-Pacific Journal of Information Technology and Multimedia, vol. 14, no. 2, pp. 472-495, 2025, doi: 10.17576/apjitm-2025-1402-09.

[31] I. Magnusson, N. A. Smith and J. Dodge, "Reproducibility in NLP: What have we learned from the checklist?," in Findings of the Association for Computational Linguistics: ACL 2023, 2023. [Online]. Available: https://aclanthology.org/2023.findings-acl.809/

Web System to Optimize Information Management for Beca 18 Selection-Stage Applicants, 2025

Jack Quispe-Vivas[1], Cristian Sedano-Contreras[2], Junior Segovia-Chalco[3], Brian Meneses-Claudio[4]
Faculty of Business Sciences, Universidad Científica del Sur, Lima, Peru[1, 2, 3]
Faculty of Business, Universidad Tecnológica del Perú, Lima, Peru[4]

***Abstract*—The objective of this study was to determine whether the implementation of a web system optimizes information management by reducing the time required to access, verify, and interpret selection-stage information for applicants of the Beca 18 2025 scholarship call. The research followed an applied, quantitative, pre-experimental design with pre-test and post-test measurements in a single group. The sample comprised 361 preselected applicants from the ordinary modality in the Lima region, selected through simple random sampling. A direct observation sheet recorded the time, in minutes, required by each participant to complete 15 tasks associated with information accessibility, information currency, and information quality before and after using the web system. The mean information-management time decreased from 130.40 to 50.46 minutes, with a statistically significant mean reduction of 79.94 minutes. The three evaluated dimensions also showed substantial reductions: information accessibility decreased by 70.3 per cent, information currency by 80.2 per cent, and information quality by 52.2 per cent. The findings indicate that, within the scope of time-based task performance, the web system optimized the efficiency with which applicants accessed, verified, and interpreted selection-stage information.**

Keywords*—Web system; information management; Beca 18; information accessibility; information currency; information quality; pre-experimental design

I. Introduction

Information management represents a persistent challenge for institutions because information is often distributed across poorly organized files, duplicated documents, or outdated versions. These conditions affect the speed with which users locate relevant content and limit the reliability of the information used for decision-making. In contexts where users must make timely decisions based on official documents, traditional or disconnected information channels can increase the effort required to access, verify, and interpret critical information.

At the global level, this problem has been widely recognized. According to a global study by the Association for Intelligent Information Management (AIIM), which compared a decade of digital-transformation efforts in organizations across North America, Europe, the Middle East, and Africa, many companies still face serious deficiencies in managing their information: 33.33 per cent of surveyed companies had not yet made notable progress in their digital-transformation processes, and more than 50 per cent continued to use paper documents and manual management [1].

At the regional level, similar challenges have been reported in organizational data management. Dell Technologies reported that 95 per cent of organizations face challenges in data management, based on 6,600 respondents from public- and private-sector organizations with more than 100 employees across Latin America, North America, Europe, Asia-Pacific, and China [2].

In Peru, the National Institute of Statistics and Informatics (INEI) found, through the 2021 Annual Economic Survey, which collected information from the 2020 economic year, that 25.9 per cent of companies did not have a management system in place [3]. This shows that a significant proportion of organizations still do not use technological tools to organize, store, or process information systematically.

In the context of the National Scholarship and Educational Loan Program (Pronabec), Beca 18 applicants must consult official information related to requirements, procedures, schedules, eligible higher education institutions, and scoring criteria during the selection stage. For the Beca 18 2025 scholarship call, this information is mainly provided through extensive official documents, which can make rapid access and interpretation difficult for applicants who need to locate specific information in a timely manner [4].

This study proposes and evaluates a web system that centralizes and structures Beca 18 selection-stage information into five interactive sections: application requirements, application procedure, schedule, eligible higher education institutions, and scoring criteria. The system was designed to reduce the time applicants spend accessing, verifying, and interpreting this information.

The general objective was to determine whether the implementation of a web system optimizes information management by reducing the time required to access, verify, and interpret selection-stage information for applicants of the Beca 18 2025 scholarship call. The specific objectives addressed three dimensions of the construct: information accessibility, information currency, and information quality.

The remainder of this study is organized as follows: Section II reviews related work and the theoretical foundations of web systems and information management. Section III describes the materials and methods, including the research design, population and sample, instrument, technological implementation, and data-analysis strategy. Section IV reports the descriptive and inferential results. Section V discusses the

findings in relation to prior studies and acknowledges the study's limitations. Section VI presents the conclusions.

II. Literature Review

A. Related Work

Prior studies have shown that web systems can improve institutional information management by reducing processing time, centralizing information, and supporting more efficient consultation processes. In an organizational context, Colan [5] implemented a web platform in a transportation company in Lima and reported a reduction in information-management time from 6.42 to 2.42 minutes, with statistically significant improvements in agility, reliability, and utility. Similarly, Alvarado [6] implemented a web system for IT-incident management at UGEL Huamanga, where the mean resolution time decreased from 59.44 to 19.11 minutes.

Other studies have reported improvements not only in time reduction but also in information registration, consultation, and reporting. Lopez [7] proposed a web application for occupational safety and health information management, increasing registration from 29.4 per cent to 94.1 per cent, consultation from 35.3 per cent to 88.2 per cent, and report generation from 23.5 per cent to 94.1 per cent. In the educational context, Chaupez and Rojas [8] reported that a web system reduced attendance-processing time from 42 to 11 seconds. Expósito and Gamboa [9] developed a web page to support information management in an educational direction system, reporting improvements in organization, task monitoring, and decision-making processes. Pinedo [10] also found statistically significant improvements across evaluated dimensions of hospital administrative information management.

Recent peer-reviewed studies provide additional evidence on the effect of information and web systems in institutional and educational contexts. Aranda-Manchay and García-Estrella [11] implemented an information system for document management in an engineering faculty using a pre-experimental design and reported a statistically significant improvement after the intervention, with $p = 0.000$ and 80 percent user acceptance. In a closer educational context, Huaman-Yupanqui et al. [12] implemented a web information system to improve academic management at UNAMAD and reported significant reductions in process times, including user registration and certificate issuance, with a Student's t-test result of $t = -78.32$ and $p < 0.001$. From a broader information-management perspective, Șandor [13] showed that web-based information management systems commonly incorporate search engines and user-management features to support information storage, retrieval, and interaction with users.

These studies support the use of web systems as tools for improving information-management processes in different institutional contexts. However, most prior work has focused on organizational, administrative, health, or academic management environments. In contrast, the present study evaluates a web system designed for applicants of a public scholarship program, focusing specifically on the time required to access, verify, and interpret official selection-stage information.

B. Theoretical Framework

A web system is a software application accessible through a browser via the internet or an intranet, allowing users to interact with services hosted on a remote server [14]. In this study, the web system functions as a technological tool for centralizing, organizing, and presenting official Beca 18 selection-stage information in a structured digital environment. This is relevant because web-based solutions can improve access to information and support decision-making when they reduce response times and provide users with clearer and more reliable content [15].

From a design perspective, usability, web accessibility, and interactivity are relevant principles for web systems because they influence how users navigate, understand, and interact with digital information environments [16], [17], [18]. In this study, these principles informed the design of the web intervention; however, the empirical evaluation focused on the dependent variable, information-management efficiency, measured through task-completion time across information accessibility, information currency, and information quality.

Information management is understood as a process that provides the resources needed for decision-making and for improving organizational processes, products, and services [19]. Effective information management also supports more accurate decision-making by ensuring that users rely on useful and reliable data [20]. In the context of this study, information management refers to the applicant's ability to access, verify, and interpret official information required during the Beca 18 2025 selection stage. Therefore, the construct was operationalized through three dimensions: information accessibility, information currency, and information quality.

Information accessibility refers to the possibility of accessing, understanding, and using available content without physical, sensory, or technological barriers. In digital environments, accessible web interfaces are essential to ensure effective access to information [21]. In this study, accessibility is associated with the time required by applicants to locate key information such as requirements, procedures, schedules, eligible higher education institutions, and scoring criteria.

Information currency refers to the degree to which available information remains updated, reliable, and useful in changing contexts. This dimension is important because official scholarship information may involve deadlines, requirements, eligible institutions, and scoring rules that applicants must verify before making decisions [22]. In this study, information currency is associated with the time required to identify whether the consulted information is current and corresponds to the valid version of the selection-stage content.

Information quality refers to the accuracy, reliability, and absence of uncertainty in data, allowing users to interpret and apply information more safely and effectively [23]. In this study, information quality is associated with the time required by applicants to interpret the meaning of requirements, procedures, schedules, eligible institutions, and scoring criteria. Thus, the three dimensions provide the conceptual basis for

evaluating the effect of the web system on information-management efficiency.

III. Materials and Methods

The study followed a sequential workflow that began with problem identification and process analysis, continued with the design and implementation of the web system, and concluded with pre-test/post-test measurement and statistical analysis. Fig. 1 presents the overall research workflow.

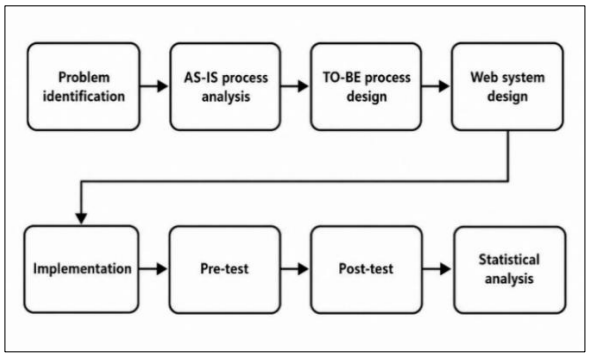


Fig. 1. Overall research workflow

A. Research Design and Study Workflow

The study was applied, quantitative, and pre-experimental. Applied research seeks to address specific problems through the use of scientific knowledge in a practical context [24]. The study followed a single-group pre-test/post-test design, O1 -> X -> O2, where O1 is the pre-test measurement, X is the web-system intervention, and O2 is the post-test measurement [25]. The explanatory scope sought to assess the changes associated with the web-system intervention in information-management outcomes; the implications of the absence of a control group are addressed in Section V. Fig. 2 summarizes the pre-experimental study process, including the sample and instrument, pre-test measurement, web-system intervention, post-test measurement, and statistical analysis.

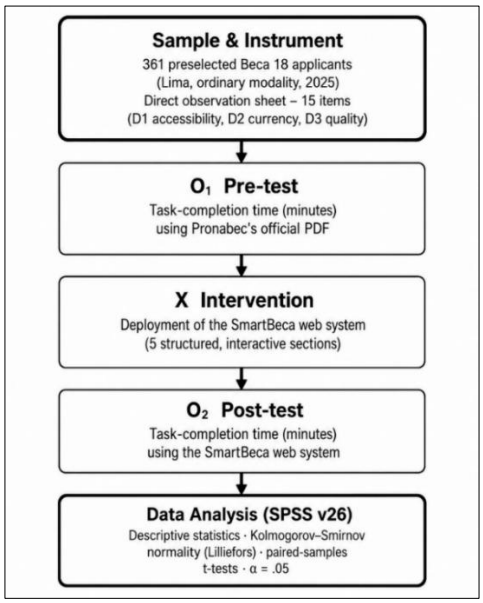


Fig. 2. Overall study process (pre-experimental single-group design).

B. Population and Sample

The target population comprised all pre-selected applicants from the ordinary modality of the Lima region for the Beca 18 2025 scholarship call. According to Pronabec [4], this population consisted of 5,830 applicants. The sample size was calculated using the finite-population formula:

$$n = (Z^2 \cdot p \cdot q \cdot N) / (e^2 (N - 1) + Z^2 \cdot p \cdot q) \quad (1)$$

With Z = 1.96, p = q = 0.5, e = 0.05, and N = 5,830, the resulting sample was 361 participants, selected through simple random sampling. Inclusion criteria required participants to be preselected applicants from the ordinary modality, from the Lima region, and from the Beca 18 2025 scholarship call.

C. Instrument

The instrument was a direct observation sheet that recorded, in minutes, the time each participant spent completing 15 observable tasks across three dimensions: information accessibility, information currency, and information quality. Times were recorded in two moments: the pre-test, in which participants used Pronabec's official PDF documentation, and the post-test, in which participants used the web system. Table I summarizes the structure of the observation instrument.

TABLE I. Observation Instrument Structure

Dimension	Items	Measurement focus
D1 — Information accessibility	1–5	Time required to locate requirements, procedures, schedule, eligible higher education institutions, and scoring criteria.
D2 — Information currency	6–10	Time required to identify whether requirements, procedures, schedule, eligible higher education institutions, and scoring criteria were current.
D3 — Information quality	11–15	Time required to interpret requirements, procedures, schedule, eligible higher education institutions, and scoring criteria.

Content validity was established through expert review by four specialists with complementary professional backgrounds in cybersecurity and information management, systems engineering, IT-project design, and public or industrial management. Because the dependent variable was operationalized as task-completion time, the instrument captures the efficiency with which applicants accessed, verified, and interpreted the required information. Complementary measures of accuracy and comprehension are discussed as a limitation in Section V.

D. Technological Implementation

The web system was built on a three-tier architecture. The frontend used Next.js, a React-based framework, to support hybrid rendering and an optimized user experience. Data persistence relied on MongoDB, a non-relational document-oriented database that provides flexible modeling and scalability. The system was deployed on Vercel, with file storage on Supabase Storage and version control through GitHub. The domain smartbeca.com was registered through Hostinger for public access, and the administrative interfaces were built with Payload CMS for structured content management. Fig. 3 shows the applicant dashboard, which centralizes the five selection-stage information sections in a single structured interface.

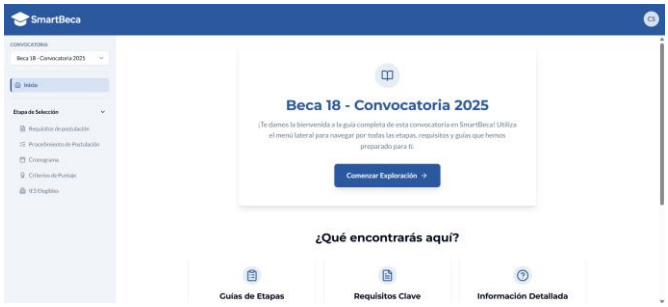


Fig. 3. SmartBeca applicant dashboard centralizes the five selection-stage information sections.

E. Data Analysis Strategy

Descriptive statistics, including mean, standard deviation, variance, minimum, and maximum values, were computed for both pre-test and post-test measurements. Normality was assessed using the Kolmogorov-Smirnov test with Lilliefors correction because the sample size was greater than 50 participants. Since normality was confirmed, paired-samples t-tests were applied to compare pre-test and post-test means for the dependent variable and each dimension. Statistical analysis was performed in SPSS v26, with the significance threshold set at $\alpha = .05$.

IV. Results

A. Descriptive Statistics

Table II presents the descriptive statistics for each dimension across both measurements. Substantial reductions were observed across all dimensions following web-system deployment. Overall information-management time decreased from a mean of 130.40 minutes in the pre-test to 50.46 minutes in the post-test, representing a 61.3 per cent reduction. The standard deviation also decreased from 11.707 to 4.216 minutes, indicating greater consistency in post-test performance.

TABLE II. Descriptive Statistics By Dimension

Dimension	*N*	*Mean*[a]	*SD*	*Var.*	*Min*	*Max*
Pre-test						
D1 — Accessibility	361	15.996	1.168	1.365	12.66	19.65
D2 — Currency	361	32.035	1.612	2.597	27.76	36.39
D3 — Quality	361	82.370	11.542	133.229	50.56	120.50
Post-test						
D1 — Accessibility	361	4.755	0.816	0.666	2.13	6.84
D2 — Currency	361	6.341	0.932	0.869	3.79	8.84
D3 — Quality	361	39.360	4.098	16.790	30.24	52.59

[a] All times are reported in minutes.

B. Normality Testing

Table III reports the Kolmogorov-Smirnov normality test results with Lilliefors correction. All eight distributions, corresponding to the three dimensions and the dependent variable in both pre-test and post-test measurements, yielded significance values above .05. Therefore, the normality assumption was met, supporting the use of parametric tests.

TABLE III. Kolmogorov-Smirnov Normality Test (Lilliefors)

Variable / Measurement	*Statistic*	*df*	*Sig. (p)*
D1 Pre-test[b]	0.033	361	0.200[a]
D1 Post-test[b]	0.032	361	0.200[a]
D2 Pre-test[b]	0.026	361	0.200[a]
D2 Post-test[b]	0.027	361	0.200[a]
D3 Pre-test[b]	0.043	361	0.098
D3 Post-test[b]	0.044	361	0.096
DV Pre-test[c]	0.044	361	0.096
DV Post-test[c]	0.037	361	0.200[a]

p-value lower-bounded at 0.200 by SPSS (Lilliefors correction)
D1, D2, D3 = dimensions of information management (Accessibility, Currency, Quality)
DV = dependent variable (Information Management)

C. Hypothesis Testing

Table IV summarizes the paired-samples t-test results. All four comparisons yielded significance values below .001, indicating statistically significant reductions in the dependent variable and the three evaluated dimensions after the web-system intervention.

TABLE IV. Paired-Samples T-Test Results (Pre-Test – Post-Test)

Comparison	*Mean Diff*[a]	*SD*	*SE*	*95% CI*	*t*	*Sig.*
DV: Information Management	79.945	12.869	0.677	78.613–81.277	118.035	<.001
D1: Accessibility	11.240	1.385	0.073	11.097–11.384	154.221	<.001
D2: Currency	25.694	1.900	0.100	25.497–25.890	256.951	<.001
D3: Quality	43.011	12.649	0.666	41.701–44.320	64.608	<.001

Mean difference reported in minutes (pre-test minus post-test); positive values indicate reduction.

For the dependent variable, overall information management, the mean difference was 79.945 minutes, with a 95 per cent confidence interval from 78.613 to 81.277 minutes. For D1, information accessibility, the mean difference was 11.240 minutes, representing a 70.3 per cent reduction. For D2, information currency, the mean difference was 25.694 minutes, representing an 80.2 per cent reduction, the largest proportional improvement. For D3, information quality, the mean difference was 43.011 minutes, representing a 52.2 per cent reduction. These results provide statistical evidence that the web system significantly reduced the time required to access, verify, and interpret selection-stage information.

V. Discussion

The findings show statistically significant reductions in the time required to complete information-management tasks after the implementation of the web system. The overall mean time decreased from 130.40 to 50.46 minutes, which is consistent with Lopez [7], who reported improvements in information registration, consultation, and report generation after deploying a web system in an occupational-safety context. Both studies

suggest that structured web interfaces can reduce the effort required to locate and use institutionally held information.

The largest relative improvement was recorded in information currency, with an 80.2 per cent reduction. Before the intervention, applicants invested considerable time verifying whether the consulted information was current, a challenge associated with static PDF-based delivery and the difficulty of identifying updated versions. This finding is consistent with Pinedo [10], whose web application for hospital administrative information yielded significant improvements in data update management. It also relates to Colan [5], who reported gains in the timeliness of information availability, a concept closely associated with information currency.

For information accessibility, which showed a 70.3 per cent reduction, the results are consistent with Alvarado [6], who reported a reduction in incident-resolution time from 59.44 to 19.11 minutes in a comparable pre-experimental design. They are also aligned with Colan [5], who found improvements in agility during information retrieval. Across these studies, centralized and structured web interfaces appear to reduce barriers that impede direct access to required information.

The information-quality dimension showed a 52.2 per cent reduction and the highest individual variability in the pre-test. This variability may be associated with differences in applicants' digital literacy and prior familiarity with the Beca 18 selection process. The result is consistent with Chaupez and Rojas [8], who reported heterogeneous improvements in an academic information-management context after the implementation of a web system. In the present study, the organized and structured presentation of information appears to have reduced the time required to interpret requirements, procedures, schedules, eligible institutions, and scoring criteria.

Several limitations qualify these findings. First, the dependent variable was operationalized solely as task-completion time. While this captures the efficiency of information retrieval and interpretation tasks, it does not establish whether applicants retrieved accurate information or interpreted it correctly. Therefore, the construct of information-management optimization is evidenced within the scope of time-based task performance, and future studies should incorporate accuracy, comprehension, and user-satisfaction measures alongside time.

Second, the single-group pre-experimental design included no control group, so threats to internal validity cannot be fully ruled out. In particular, the Hawthorne effect, the testing effect, and maturation are plausible partial explanations for the observed reductions, especially for tasks requiring longer interpretation time. Quasi-experimental designs with matched control groups and counterbalanced task orders would strengthen causal attribution in future research.

Third, the sample was restricted to ordinary-modality applicants in the Lima region, so generalization to other regions or modalities warrants caution. Rural areas of Peru face heterogeneous connectivity and device availability, which could affect both system adoption and the magnitude of the time savings reported here. Evaluating the system under diverse connectivity conditions, including low-bandwidth and mobile-first scenarios, is a relevant direction for future work and would broaden the practical interest of the solution.

VI. Conclusion

This study showed that implementing the web system significantly improved the efficiency of information management for preselected applicants of the Beca 18 2025 scholarship call. Overall mean time decreased from 130.40 to 50.46 minutes, supporting the general hypothesis. The system functioned as a tool for centralizing, organizing, and presenting selection-stage information, enabling applicants to access, verify, and interpret it in less time and with greater autonomy.

Information currency recorded the greatest relative improvement, indicating that the permanent availability of updated and centralized content is a relevant factor in information-management efficiency within high-stakes application contexts. Information accessibility also improved substantially, showing that structured web navigation helped reduce the barriers associated with PDF-based information delivery. Information quality improved as well, indicating that the way a system organizes and presents information can reduce the time required for users to interpret and apply it.

The findings support recommending that Pronabec adopt web-based information platforms as a primary information channel for future Beca 18 calls, with protocols to ensure periodic content updates. Future studies should replicate the research using quasi-experimental designs with control groups and incorporate complementary measures such as accuracy, comprehension, and user satisfaction to strengthen the evidence base for this type of intervention in educational-inclusion programs.

Acknowledgment

We thank the Universidad Científica del Sur for the support and training provided during the development of this research, as well as the preselected Beca 18 applicants who participated in the study and shared their experiences. As former beneficiaries of the program, we hope this work contributes to clearer and more timely access to scholarship information for future applicants.

References

[1] Association for Intelligent Information Management (AIIM), AIIM 2023 State of the Intelligent Information Management Industry: Pivotal Moment in Information Management, 2023. [Online]. Available: https://www.opentext.com/assets/documents/en-US/pdf/opentext-report-aiim-2023-state-of-the-intelligent-information-management-industry-en.pdf

[2] Dell Technologies, Innovation Catalysts Study, 2024

[3] Instituto Nacional de Estadística e Informática (INEI), Perú: Tecnologías de Información y Comunicación en las Empresas, 2020. Lima, Perú, 2025.

[4] Pronabec, Expediente técnico del concurso Beca 18 2025. Lima, Perú, 2025. [Online]. Available: https://cdn.www.gob.pe/uploads/document/file/6938514/5987321-rde-n-113-2024-minedu-vmgi-pronabec.pdf

[5] J. Colan, "Implementación de un sistema web para optimizar la gestión de información en una empresa de transporte ubicada en Lima, Perú 2021," B.S. thesis, Universidad Tecnológica del Perú, Lima, 2021.

[6] J. Alvarado, "Sistema web en la gestión de incidencias de tecnologías de información en la UGEL Huamanga 2021," B.S. thesis, Universidad Continental, Huancayo, 2023.

[7] R. López, "Propuesta de implementación de un sistema web para mejorar la gestión de la información del área de seguridad y salud ocupacional de Arca Continental Lindley, Chilca-2023," B.S. thesis, Universidad Nacional José Faustino Sánchez Carrión, Huacho, 2024.

[8] J. Chaupez and A. Rojas, "Sistema web para la gestión de la información académica en la I.E.P. 20189 Nuestra Señora de la Concepción - Cañete, 2023," B.S. thesis, Universidad Nacional de Cañete, Cañete, 2023.

[9] P. Expósito and M. Gamboa, "Página web para gestionar información en el sistema de dirección educacional," Didasc@lia: Didáctica y Educación, vol. 12, no. 4, pp. 240–261, 2021.

[10] S. Pinedo, "Aplicación web para la gestión de la información del personal administrativo y asistencial del Hospital Regional Docente Daniel Alcides Carrión," B.S. thesis, Universidad Peruana de los Andes, Huancayo, 2022.

[11] H. R. Aranda-Manchay and C. W. García-Estrella, "Sistema de información para la gestión documental en una Facultad de Ingeniería," Revista Científica de Sistemas e Informática, vol. 3, no. 2, e521, 2023, doi: 10.51252/rcsi.v3i2.521.

[12] N. C. Huaman-Yupanqui, Y. Zapana-Halire, N. J. Ulloa-Gallardo, and Y. Vilchez Navarro, "Implementación de un sistema web para mejorar la gestión académica en universidades: un estudio de caso," Revista Amazonía Digital, vol. 3, no. 2, e289, 2024, doi: 10.55873/rad.v3i2.289.

[13] A. Șandor, "Web-based information management systems feature usage: Case study on European telecom sector," TEM Journal, vol. 10, no. 2, pp. 592–596, 2021, doi: 10.18421/TEM102-14.

[14] L. Mayorga, O. Bermeo, G. Riccardi, and V. Guevara, "Sistema web para los procesos administrativos y de producción en viveros del Cantón Milagro," Revista Ingeniería, vol. 6, no. 16, pp. 200-213, 2022.

[15] L. Silva and W. Farro, "Desarrollo de un sistema web basado en arquitectura REST API para la toma de decisiones en un e-commerce en gastronomía," B.S. thesis, Universidad Tecnológica del Perú, Lima, 2024.

[16] I. Condo, "Análisis, diseño y maquetación de un sitio web para la empresa Calzado Ale partiendo de la utilización de una metodología para cumplir parámetros de usabilidad," B.S. thesis, Universidad Nacional de Chimborazo, Riobamba, 2024.

[17] C. Esteve and S. Mascarell, Guía básica de accesibilidad web. Sendemà Editorial, 2023. [Online]. Available: https://dicatic.webs.upv.es/wp-content/uploads/2024/10/Guia-basica-para-crear-documentos-accesibles-WEB.pdf

[18] K. Cerna, K. Duran, W. Genovez, and F. Aguilar, "El beneficio del uso de las herramientas web en el sector educativo," Revista Arbitrada Interdisciplinaria Koinonía, vol. 8, no. 2, pp. 234–251, 2024, doi: 10.35381/r.k.v8i2.2874.

[19] K. M. Domínguez Rodríguez, L. Carralero Hidalgo, and M. A. Tamayo Fajardo, "La gestión del sistema de información y su concepción en la proyección estratégica del posgrado universitario," in Desarrollo social con calidad de vida para América Latina, J. Álvarez Botello, E. M. Chaparro Salinas, J. A. Ruiz Tapia, C. E. Estrada Gutiérrez, M. del C. Hernández Silva, and M. de la L. I. Sánchez Paz, Comps. Toluca, México: Universidad Autónoma del Estado de México, 2023, pp. 292-305.

[20] M. E. Gamboa-Graus, Y. Castillo-Rojas, and J. F. Parra-Rodríguez, "Procedimiento para la gestión de información en función de la administración escolar," Dilemas Contemporáneos: Educación, Política y Valores, vol. 8, no. 3, 2021, doi: 10.46377/dilemas.v8i3.2625.

[21] M. Chamorro, A. Duarte, N. Calderón, S. Duarte, and V. Jiménez, "Accesibilidad de información en la web de instituciones oficiales de Paraguay," Palabra Clave (La Plata), vol. 10, no. 2, pp. 1-8, 2021, doi: 10.24215/18539912e129.

[22] J. Pillihuamán, "Percepción de la gestión de información en una subprefectura distrital de Perú," Revista de Ciencia e Investigación en Defensa - CAEN, vol. 6, no. 1, pp. 61-74, 2025.

[23] M. Vergara, "Descripción de estrategias para mejorar la calidad de información en órdenes de servicio energético," Inventum, vol. 16, no. 30, pp. 91-102, 2021, doi: 10.26620/uniminuto.inventum.16.30.2021.91-102.

[24] J. Castro-Maldonado, L. Gómez-Macho, and E. Camargo-Casallas, "La investigación aplicada y el desarrollo experimental en el fortalecimiento de las competencias de la sociedad del siglo XXI," Tecnura, vol. 27, no. 75, pp. 140–174, 2022, doi: 10.14483/22487638.19171.

[25] C. A. Ramos, "Diseños de investigación experimental," CienciAmérica, vol. 10, no. 1, pp. 1-7, 2021, doi: 10.33210/ca.v10i1.356.

A Modified Lyapunov-Based MEC Offloading Algorithm for Severity-Aware QoS in 5G/B5G IoT Systems

Modified Lyapunov-Based Severity-Aware QoS in 5G/B5G IoT Systems

Sahana S Reddy*, R. Sukumar
Department of Electronics and Communication Engineering, JAIN (Deemed to be University), Bengaluru, 560 034, India

***Abstract*—The rapid growth of latency-sensitive and computation-intensive IoT applications in 5G and Beyond-5G (B5G) networks has increased the demand for efficient Multi-access Edge Computing (MEC) offloading strategies. Current MEC frameworks have several limitations: 1) binary QoS modeling without considering deadline violation severity, 2) a lack of severity-aware optimization in IoT applications, 3) insufficient consideration of different task criticality, and 4) poor handling of dynamic latency and energy trade-off in large-scale IoT environments. This study proposes a modified Lyapunov-based severity-aware MEC offloading framework for heterogeneous 5G/B5G IoT systems. The proposed framework utilizes task deadlines, task criticality, queue states, wireless channel conditions, and MEC resource availability as input for adaptive offloading optimization. A QoS Violation Severity Index is introduced to jointly capture deadline violation magnitude and task criticality. Furthermore, severity-aware virtual queues are integrated with a modified Lyapunov Drift-Plus-Penalty optimization framework to dynamically minimize QoS violation severity while balancing latency and energy consumption. Experimental evaluation demonstrates that the proposed framework significantly reduces average task delay to 82 ms, energy consumption to 6.0 mJ, and QoS violation rate to 4.8%, while improving long-term system stability compared with existing MEC offloading approaches in dynamic 5G/B5G IoT environments.**

***Keywords*—5G and Beyond-5G; multi-access edge computing; quality of service; task offloading; severity estimation; lyapunov optimization**

I. Introduction

The evolution of 5G and Beyond-5G (B5G) networks have greatly improved the capabilities of the Internet of Things (IoT) by facilitating ultra-reliable communication, massive device connectivity and low-latency services [1]. These networks enable real-time applications like smart healthcare, autonomous transportation, industrial automation, intelligent surveillance, and smart city infrastructures, which produce and process huge amounts of heterogeneous data [2, 3]. To facilitate such latency-sensitive applications, Multi-access Edge Computing (MEC) has been embedded into 5G/B5G architectures to introduce computation and storage to the end-users and thereby shorten the communication latency and improve service responsiveness [4, 5]. Although the 5G/B5G IoT technologies have made great strides, modern network has still many challenges in operation and resource management [6]. The massive growth of connected devices introduces network congestion, complexity of resource allocation, and changes in traffic patterns, making the efficient management of tasks highly challenging [7, 8]. Moreover, IoT devices also have limited battery power, computation capability, and storage capacity, which increases the reliance on edge servers [9, 10]. Furthermore, heterogeneous IoT applications show a wide variety of Quality of Service (QoS) demands related to latency, reliability, bandwidth, and energy efficiency, making uniform resource management ineffective [11, 12]. Therefore, achieving reliable QoS under various latency, energy consumption, and resource utilization constraints remains a significant challenge in the modern 5G and B5G IoT environments [13].

Several existing studies analyze MEC-based task offloading and resource optimization techniques to improve QoS performance in 5G and B5G IoT networks [14, 15]. For instance, a joint optimization framework for energy consumption and time delay in dependency-aware task offloading for 5G applications was developed [16]. It used a Directed Acyclic Graph (DAG) to analyze the dependencies between tasks and an improved Particle Swarm Optimization (PSO) algorithm to optimize task offloading decision and MEC server selection optimization. To efficiently manage the task, a 5G-MEC task offloading framework was developed, which minimizes dropped task ratio, computational latency, and communication latency through Mixed Integer Linear Programming (MILP), PSO, and Genetic Algorithm (GA) [17]. Despite the progress, several key research gaps remain. 1) Existing works usually model QoS violations in a binary way without considering the severity of deadline misses, making it ineffective in handling latency-sensitive IoT applications [18,19]; 2) Most existing schemes focus on optimizing average latency or energy consumption rather than explicitly minimizing severe QoS violations, which expose critical tasks to risks [20]; 3) Current MEC offloading frameworks consider limited heterogeneous task criticality, leading to unfair QoS degradation for high-priority applications [21, 22]; 4) Many existing solutions assume relatively homogeneous IoT environments and fail to effectively address the diverse latency, reliability, and energy requirements of large-scale heterogeneous IoT applications in dynamic 5G and B5G networks [23, 24]. To overcome these, the proposed work develops a severity-aware MEC task offloading framework for 5G and Beyond-5G IoT networks that minimizes

*Corresponding author.

severe QoS violations while balancing latency and energy consumption for heterogeneous IoT applications. To achieve this, a modified Lyapunov Drift-Plus-Penalty (LDPP) optimization framework is developed to enable dynamic and real-time offloading decisions by jointly stabilizing severity-aware virtual queues and optimizing system performance under varying network conditions. The key novelty of the proposed framework is as follows:

Severity-Aware QoS Violation Modeling: A novel QoS Violation Severity Index (QVSI) is proposed to quantify the severity of the delay violation by combining the task criticality and the magnitude of delay violation, overcoming the limitations of traditional binary QoS models in 5G/B5G IoT networks.

Modified Lyapunov-Based Optimization Framework: A severity-aware virtual queue model is incorporated into a modified LDPP optimization framework to jointly minimize the severity of QoS violation, LATENCY, and energy consumption while ensuring long-term stability.

Adaptive MEC Offloading Decision Strategy: An intelligent offloading mechanism is developed to dynamically select local execution, MEC offloading, and cooperative MEC execution based on the real-time queue states, channel conditions, and the available MEC resources in the large-scale heterogeneous IoT environment.

Efficient Support for Heterogeneous 5G/B5G IoT Applications: The proposed framework effectively handles heterogeneous IoT applications with different latency requirements and criticality levels, while minimizing the occurrence of severe QoS violations, deadline miss ratio, latency, and energy consumption compared with existing MEC offloading approaches.

The article is structured as follows. A review of the relevant works is presented in Section II. Section III includes a detailed process of the proposed approach for task offloading in MEC-enabled IoT systems. Section IV exhibits the findings and discussion, and Section V concludes with the findings of the study.

II. Related Work

This section reviews the related work on existing MEC-based task offloading and resource allocation methods in 5G/B5G IoT environments. Zhai et al. [25] modeled a task-offloading scheme for hybrid edge computing networks, where the tasks can be executed locally, offloaded to edge servers, or forwarded to neighboring outgoing devices. It used Deep Reinforcement Learning (DRL) with a Double Deep Q Network (DDQN) algorithm to optimize the task offloading decisions, reducing the overall task delay and prioritizing the high-priority tasks. However, the framework mainly focused on delay-aware offloading and priority scheduling, whereas the severity of QoS violations, heterogeneous task criticality, and long-term queue stability were not explicitly modeled in the optimization process. Benbraika et al. [26] introduced the Hungarian Algorithm for Task Offloading (HATO) for effective task scheduling in 5G-enabled Vehicular Edge Computing (VEC) systems. The model effectively distributed computational workloads, reduced delay, and improved task processing performance in autonomous vehicles. However, the approach mainly depended on stable 5G connectivity and central edge server support, which may affect performance in highly dynamic real-time environments.

To balance the trade-off between energy consumption and computation time, Alam et al. [27] introduced task offloading and resource allocation for the Internet of Vehicles (IoV) using PSO. The method improved QoS, reduced latency, and minimized energy consumption during computational task execution. However, the model was estimated under limited network settings, and it could not capture task criticality during offloading optimization, which limits its applicability to delay-sensitive applications. Younis et al. [28] introduced an Energy-Latency-aware Task Offloading and Approximate Computing (ETORS) framework to optimize the trade-off between energy consumption and latency using the Dual-Decomposition Method (DDM). This framework intelligently balances the energy consumption and shortened completion time by processing the workloads without overloading. However, the framework primarily focused on energy-latency trade-off optimization without considering severity-aware QoS degradation and long-term queue stability. Luo et al. [29] introduced a joint task offloading and resource allocation framework for MEC using the Lyapunov optimization with the Potential Minimum Point (PMP) algorithm. The model achieved increased QoS, lowered latency, balanced energy consumption, and better system stability. However, the framework focused mainly on resource optimization and system stability, while the effect of deadline violation and differentiated QoS handling for critical IoT tasks remains unexplored. Yang et al. [30] introduced a Deep Reinforcement Learning-based Task Offloading (DRTO) framework in MEC systems using Deep Neural Networks (DNNs) and Reinforcement Learning (RL) techniques. The method helped reduce task delay, improved computation efficiency, and increased the speed of offloading decisions under changing wireless network conditions. However, the framework mainly focused on delay optimization and could not jointly optimize QoS violation severity, latency, and energy consumption under long-term queue stability constraints in heterogeneous IoT environments.

Mathi et al. [31] introduced a Distributed Collaborative Learning Framework to detect Black Hole Attacks in the Routing Protocol for Low-power and Lossy networks (RPL) based Internet of Things (IoT) networks. The framework reduces communication overhead, energy consumption, and computational complexity. However, the framework is mainly focused on detecting black hole attacks; performance in IoT deployments cannot be discussed. Xiao and Dong [32] introduced a lightweight CNN-based deep learning model for graphic recognition in the Internet of Things (IoT) ecosystem. The model was suitable for resource-limited IoT devices and supports simultaneous graphic recognition applications. However, the impact of varying network conditions and device heterogeneity was not analyzed. Vidyaningtyas et al. [33] introduced a Sequential Power Allocation (SePA) algorithm to optimize power distribution within each sector. The method minimizes intra-sector interference and improves power efficiency, and the framework supports reliable communication in high-density environments such as Internet of Things (IoT)

networks. However, the machine learning-based optimization techniques are not integrated.

Thus, by analyzing extensive research on MEC-based task offloading and resource optimization in 5G and B5G IoT networks. Existing approaches mainly focus on minimizing the average latency, energy consumption, or overall task failure rate without explicitly considering the severity of QoS violations. Most existing approaches consider QoS satisfaction in a binary manner and ignore the extent to which deadlines are missed, which is critical for latency-sensitive and mission-critical IoT applications. Moreover, there is a limited consideration of heterogeneous task criticality and diverse QoS requirements, leading to suboptimal and sometimes unfair resource allocation for high-priority tasks. As a result, current MEC offloading frameworks remain insufficient in effectively handling severe QoS degradation under dynamic and large-scale IoT network conditions. Therefore, the proposed framework develops a severity-aware LDPP-based offloading algorithm that minimizes severe QoS violations while balancing latency and energy consumption.

III. System Overview

The proposed system considers a 5G/B5G-enabled MEC architecture consisting of multiple heterogeneous IoT devices, a 5G/6G base station, and an edge server connected through wireless communication links, as shown in Fig. 1. Each IoT device creates computation-heavy and latency-sensitive tasks that are stored in local task queues. The IoT devices have different battery life, computing power, and QoS needs, making the network highly dynamic and resource-constrained. Depending on various wireless channel conditions, queue status, task deadline, and device energy level, each task may be processed locally on the IoT device or offloaded to the MEC/edge cloud server for faster processing. The wireless communications between the IoT devices and the MEC server are represented through time-varying channel states, which represent real-world network dynamics. To effectively provide delay-sensitive and critical IoT applications in 5G/B5G environments, the proposed framework introduces a severity-aware task management scheme that measures the QoS violation, optimizes offloading decisions adaptively to minimize the latency, energy consumption, and service reliability.

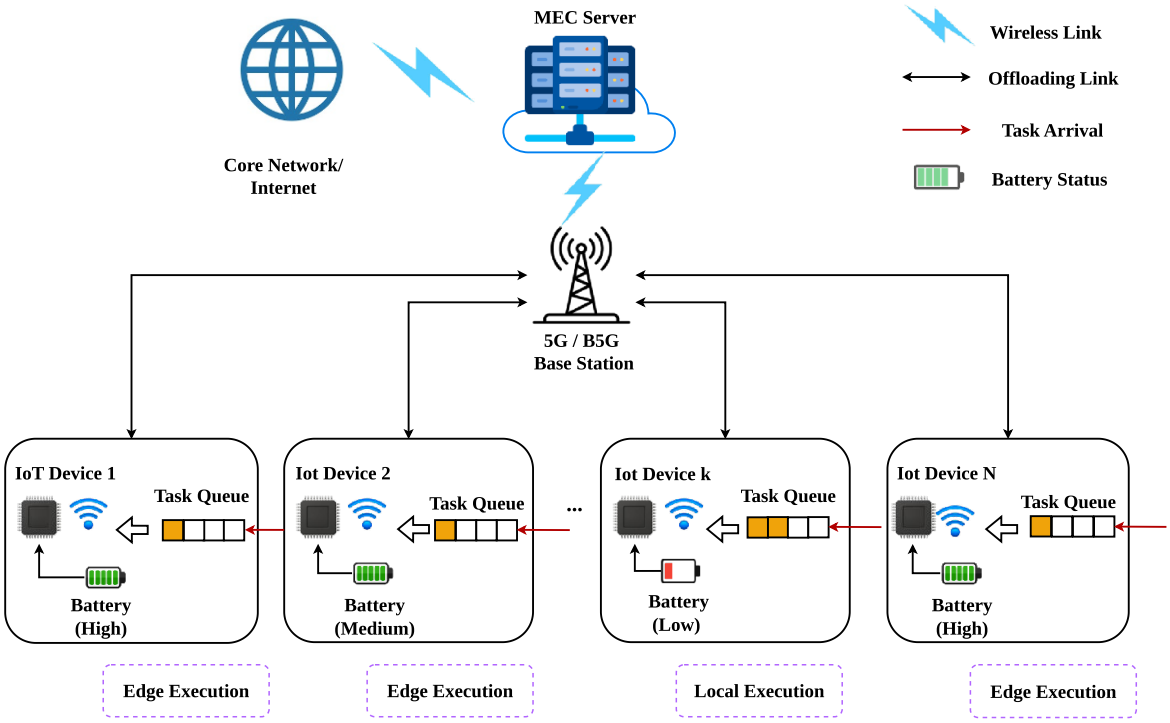


Fig. 1. System model for multi-access edge computing-assisted task offloading in heterogeneous 5G/B5G IoT environments.

A. Problem Formulation

The considered 5G/B5G MEC-enabled IoT system supports multiple heterogeneous IoT devices that generate computation-intensive and latency-sensitive tasks to be executed locally or offloaded to the MEC server. The primary goal is to reduce the QoS violation severity, latency, and energy consumption, while maintaining long-term system stability under stochastic task arrival. In contrast to traditional binary QoS models, a severity-aware optimization framework is incorporated to account for the magnitude of deadline misses and for task importance. The optimal long-term problem is formulated as shown in Eq. (1).

$$min\ lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[\sum_{i=1}^{N}\left(\alpha S_i(t) + \beta T_i(t) + \gamma E_i(t)\right)\right] \quad (1)$$

where, $\mathrm{E}[\cdot]$ representing long-term average system behavior, $S_i(t)$, $T_i(t)$, and $E_i(t)$ represent the severity of QoS violation, latency and energy consumption, respectively, N is the total number of IoT devices, T is the total simulation, and t is the discrete time slot. The weighting factors α, β, and γ are such that $\alpha + \beta + \gamma = 1$, controlling the balance between severity,

delay, and energy efficiency. System stability is guaranteed by resource constraints as the sum of computation capacity of MECs is bounded by $\sum_{i\epsilon N} f_i(t) \leq F^{max}$ and the sum of communication bandwidths of MECs is bounded by $\sum_{i\epsilon N} b_i(t) \leq B^{max}$. The components $f_i(t)$ and $b_i(t)$ are the allocated computing and bandwidth resources, and F^{max} and B^{max} are the maximum capacities of computing and bandwidth, respectively. To achieve long-term stability, the virtual queue should be bounded as in Eq. (2).

$$lim_{T\to\infty} sup \frac{1}{T} \sum_{t=1}^{T} E[Z_i(t)] < \infty \quad (2)$$

where, $Z_i(t)$ is the severity-aware queue length. Lastly, exactly one execution mode is chosen for each task, in which the decision variables are binary. The problem is formulated as a stochastic optimization problem that minimizes the severity of the long-term QoS violations, the system latency, and the system energy consumption under the system constraints.

B. Proposed Severity-Aware Offloading Framework

The proposed severity-aware MEC offloading framework sequentially and iteratively operates to efficiently manage computation tasks in 5G/B5G IoT environments, as depicted in Fig. 2. Initially, the IoT devices generate heterogeneous tasks, characterized by data size, deadline constraints, and levels of criticality. The system estimates the local execution and MEC offloading delay, including transmission and computation for each task. Based on these delays, the QVSI is calculated to reflect the degree of possible violation of the deadlines while incorporating task importance. The obtained severity values are then used to update severity-aware virtual queues, which capture the accumulated QoS degradation over time. These queue states are then embedded in an LDPP optimization framework to optimize latency, energy use, and severity-aware QoS performance. This optimization dynamically switches between local processing mode, MEC offloading mode, and cooperative MEC execution mode based on the system conditions of each task. The selected decisions are then implemented, and the resulting delay and energy consumption are recorded for performance assessment. Finally, the severity-aware queues are updated based on the observed outcomes, and the process repeats over the successive time slots, which makes real-time adaptive and stable offloading decisions under the dynamic network conditions.

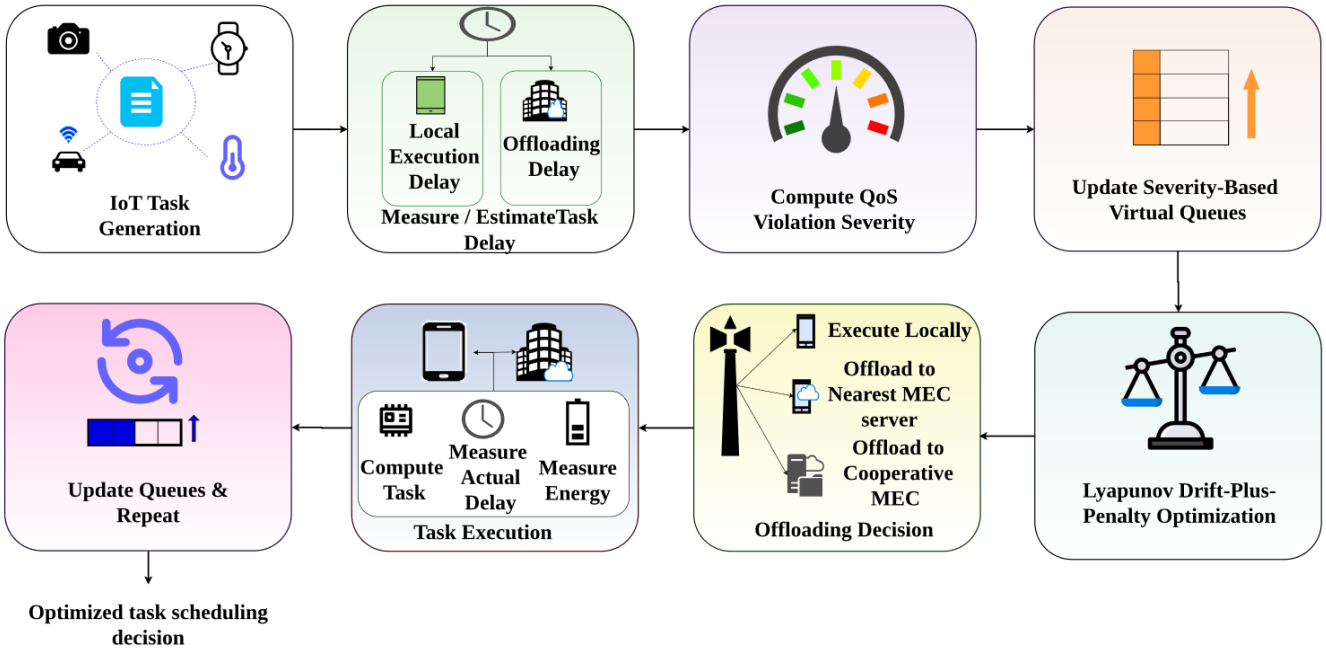


Fig. 2. Architecture of the proposed severity-aware Lyapunov-based MEC offloading system for 5G/B5G IoT networks.

1) System initialization and task modelling: In the considered 5G/B5G MEC enabled IoT environment, a collection of heterogeneous IoT devices $\mathcal{N} = \{1,2,\dots,N\}$ dynamically generates computation tasks over time. Every task generated by the i-th device $\mathcal{T}_i(t)$ is expressed as follows in Eq. (3).

$$\mathcal{T}_i(t) = \{S_i^{data}(t), D_i^{max}, C_i\} \quad (3)$$

where, $S_i^{data}(t)$ is the task size, D_i^{max} is the deadline constraint, and C_i is the task criticality. The task criticality factor C_i represents the priority level of each IoT task and is assigned based on the application type and service requirements. In the proposed framework, C_i is predefined at the time of task generation and remains static throughout the task execution process. Specifically, high-criticality tasks (e.g., URLLC-based applications) are assigned higher weights, medium-criticality tasks are assigned moderate weights, and low-criticality tasks are assigned lower weights. This static assignment reflects the inherent priority differences among heterogeneous IoT services and ensures consistent severity-aware scheduling and offloading decisions within the Lyapunov optimization framework. Each task is initially stored in the local queue of each IoT device upon arrival, and then a decision is taken to either execute or offload the task. The local task dynamics at each device are modeled as indicated in Eq. (4).

$$Q_i(t+1) = max[Q_i(t) - \mu_i(t), 0] + A_i(t) \quad (4)$$

where, $Q_i(t)$ is the queue backlog of the device i, $A_i(t)$ is the number of newly arrived tasks at the time slot t, $\mu_i(t)$ is the number of tasks processed or offloaded at the time slot t, and $Q_i(t+1)$ is the updated queue backlog of IoT devices i at the next time slot $t+1$. This queue formulation represents the stochastic network conditions with dynamic arrival and service processes of IoT tasks. The defined queueing model serves as the foundation for subsequent severity-aware optimization and Lyapunov-based offloading decisions in the proposed framework.

2) Task characterization and delay estimation: For each incoming task i, the system evaluates its execution characteristics under two different modes: local execution and MEC offloading [34]. The local execution delay $D_i^{loc}(t)$ is dependent on the computational power of the IoT device and is represented as given in Eq. (5).

$$D_i^{loc}(t) = \frac{S_i^{data}(t).k}{f_i^{loc}(t)} \quad (5)$$

where, $S_i^{data}(t)$ is the size of the task, k is the number of CPU cycles needed per bit of data, and $f_i^{loc}(t)$ is the local CPU frequency. In the case of MEC offloading, the overall delay consists of the transmission delay and the edge computation delay as indicated in Eq. (6).

$$D_i^{mec}(t) = D_i^{trans}(t) + D_i^{edge}(t) \quad (6)$$

where, $D_i^{mec}(t)$ denotes total MEC offloading delay, $D_i^{trans}(t)$ denotes uplink data transmission delay, and $D_i^{edge}(t)$ denotes the delay in data processing. The system chooses the minimum possible delay based on the network condition, channel state, and available resources. The resulting estimated delay $D_i(t)$ is referred as shown in Eq. (7).

$$D_i(t) = min\left(D_i^{loc}(t), D_i^{mec}(t)\right) \quad (7)$$

This process is subsequently applied in QoS violation severity computation and offloading optimization decisions.

3) QoS violation severity modeling: In conventional MEC offloading systems, QoS is generally considered in a binary way, either a task meets or fails its deadline constraint. However, such modeling fails to capture the extent of violation, an important aspect in 5G/B5G IoT systems supporting heterogeneous and latency-sensitive applications. To overcome this limitation, the delay violation function $V_i(t)$ [35] of task i at the time slot t is defined as shown in Eq. (8).

$$V_i(t) = max(0, D_i(t) - D_i^{max}) \quad (8)$$

where, $D_i(t)$ is the actual or estimated time to complete the task and D_i^{max} is the maximum permissible task completion time. The function guarantees that only positive deadline deviations are considered, and the penalty for non-violated tasks is zero.

To further model heterogeneity in the IoT applications, each task is given a criticality factor C_i denoting the priority level and importance of the task. By combining delay violation magnitude and the importance of the task, the QVSI [36] is defined as follows in Eq. (9).

$$S_i(t) = C_i \cdot V_i(t) \quad (9)$$

where, $S_i(t)$ is the severity of the QoS violation of the task i. The final severity model is given as follows in Eq. (10).

$$S_i(t) = C_i \cdot max(0, D_i(t) - D_i^{max}) \quad (10)$$

A QoS violation is considered to occur when the task completion delay exceeds its deadline, i.e., $D_i(t) > D_i^{max}$. However, unlike conventional binary models, the proposed framework does not treat QoS violation as a binary event. Instead, the severity of the violation is quantified using the QVSI $S_i(t)$, which jointly captures both the magnitude of deadline violation and the task criticality. Therefore, although violation detection is based on deadline exceeding, the decision-making process is entirely driven by the severity value $S_i(t)$, rather than a binary QoS indicator. This formulation makes it possible to jointly consider the degree of violation of the deadline and the criticality of the task in the system design. The proposed QVSI serves as a key input for the Lyapunov-based optimization framework by updating virtual queues according to the severity and influencing real-time offloading decisions.

4) Severity-aware virtual queue construction: To guarantee long-term QoS violation control, a severity-aware virtual queue is introduced for each IoT device. The idea behind this virtual queue is to redefine the problem of managing QoS violations as a stochastic queue stability problem, allowing the use of Lyapunov optimization for offloading decisions at real-time [37].

In particular, let $S_i(t)$ be the QVSI for the task i in the time slot t. The severity-based virtual queue evolution $Z_i(t+1)$ is defined as shown in Eq. (11).

$$Z_i(t+1) = max[Z_i(t) + S_i(t), 0] \quad (11)$$

where, $Z_i(t)$ represents the accumulated severity backlog of the device i at time t with initial condition zero. The virtual queue accumulates the QoS violation severity over time, such that the higher the value indicates severe or persistent the deadline violations. This allows prioritization of high-severity tasks during scheduling and offloading decisions.

By including $Z_i(t)$ into the Lyapunov framework, the QoS violation minimization problem is converted into a queue stability problem. This enables real-time optimization of latency, energy consumption, and severity-aware QoS performance. This formulation dynamically penalizes continually violated tasks and provides long-term stability of the system with respect to its QoS.

5) Lyapunov drift-plus-penalty formulation: The proposed severity-aware MEC offloading framework minimizes the long-term QoS violation severity by stabilizing the severity-aware virtual queues, while simultaneously optimizing latency and energy consumption through the penalty component of the Lyapunov Drift-Plus-Penalty framework [38]. The system stability is characterized using a Lyapunov function $L(t)$

defined over the severity-aware virtual queues as shown in the Eq. (12).

$$L(t) = \frac{1}{2}\sum_{i=1}^{N} Z_i^2(t) \quad (12)$$

where, N is the number of the total IoT devices and $Z_i(t)$ is the backlog of virtual queues for IoT device i at time t determined by its severity. The term $Z_i^2(t)$ helps to ensure system stability by providing a higher penalty if the queue backlog is larger.

To jointly optimize system performance and queue stability, the one-slot conditional Lyapunov drift $\Delta L(t)$, is defined as shown in Eq. (13).

$$\Delta L(t) = \mathbb{E}[L(t+1) - L(t) | Z(t)] \quad (13)$$

where, $L(t+1)$ is the Lyapunov function value at the next time slot $t+1$ and $\mathbb{E}[\cdot]$ denotes the expectation taken over the system randomness, including the arrival of tasks, channel variations, and offloading decisions.

The Lyapunov drift inherently captures the accumulated QoS violation severity through the severity-aware virtual queue $Z_i(t)$, whose evolution is governed by the QoS Violation Severity Index $S_i(t)$ in Eq. (11). Therefore, minimizing the Lyapunov drift implicitly minimizes the long-term QoS violation severity. Consequently, the penalty term is introduced only for latency and energy consumption, avoiding redundant optimization of the severity objective. The resulting Drift-Plus-Penalty (DPP) function is expressed in Eq. (14).

$$\mathcal{O}_t = \Delta L(t) + V \cdot \mathbb{E}[\lambda P(t) + \mu T(t) | Z(t)] \quad (14)$$

where, $\mathcal{O}_t$ represents the combined Lyapunov drift and penalty function, $P(t) = \sum_{i=1}^{N} P_i(t)$ represents total energy consumption at time t, and $T(t) = \sum_{i=1}^{N} T_i(t)$ represents the total task latency in the system. The parameter $V > 0$ determines the trade-off between maintaining queue stability and optimizing system performance, while λ and μ ($\lambda + \mu = 1$) control the relative importance of latency and energy consumption in the penalty term. Specifically, smaller values of V place greater emphasis on stabilizing the severity-aware virtual queues, thereby reducing long-term QoS violation severity, whereas larger values prioritize the optimization of latency and energy consumption. In this work, the value of V was selected based on preliminary simulation experiments to achieve a balanced trade-off between QoS violation severity, latency, and energy consumption. Unlike the overall optimization objective in Eq. (1), which considers QoS violation severity, latency, and energy consumption, the Drift-Plus-Penalty formulation optimizes the QoS violation severity implicitly through the Lyapunov drift via the severity-aware virtual queues, while latency and energy consumption are optimized explicitly through the penalty term. Thus, the proposed optimization remains fully consistent with the long-term objective while enabling adaptive real-time MEC offloading in 5G/B5G IoT networks. This detailed process is shown in Fig. 3.

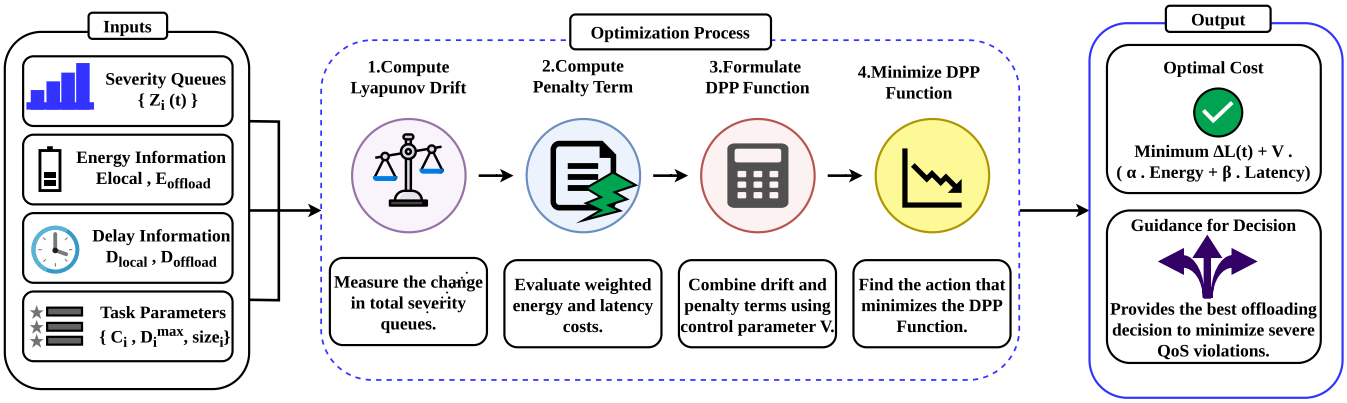


Fig. 3. Severity-aware Lyapunov Drift-Plus-Penalty (LDPP) optimization process in MEC-enabled 5G/B5G IoT networks.

Theoretical Stability Analysis: To theoretically justify the long-term stability of the proposed severity-aware MEC offloading framework, we analyze the system using Lyapunov optimization under stochastic network conditions.

The following standard assumptions are considered: 1) task arrivals follow a stochastic process with bounded first and second moments, 2) wireless channel states are time-varying and stochastic but remain bounded within a finite support, 3) task sizes, computational workloads, and delay deadlines are finite and bounded, and 4) system resources including computation and communication capacities are finite and constrained as defined in the system model.

Theorem 1 (Queue Stability Guarantee): Under the above assumptions, the proposed severity-aware Lyapunov Drift-Plus-Penalty (DPP) based offloading policy ensures that the severity-aware virtual queues $Z_i(t)$ are mean-rate stable. Furthermore, the time-average expected queue backlog remains bounded, guaranteeing long-term stability of the MEC system under stochastic task arrivals and time-varying wireless channel conditions.

Proof:

The Lyapunov function is defined in Eq. (15).

$$L(t) = \frac{1}{2}\sum_{i=1}^{N} Z_i^2(t) \quad (15)$$

The conditional Lyapunov drift is given by Eq. (16).

$$\Delta L(t) = \mathbb{E}[L(t+1) - L(t) \mid Z(t)] \quad (16)$$

Using the queue evolution, it is expressed as shown in Eq. (17).

$$Z_i(t+1) = \max[Z_i(t) + S_i(t), 0] \quad (17)$$

The drift can be upper-bounded as shown in Eq. (18).

$$\Delta L(t) \leq B + \sum_{i=1}^{N} Z_i(t)\mathbb{E}[S_i(t) \mid Z(t)] \quad (18)$$

where, B is a finite constant due to bounded severity values. The proposed DPP policy minimizes as expressed in Eq. (19).

$$\Delta L(t) + V \cdot \mathbb{E}[\lambda P(t) + \mu T(t) \mid Z(t)] \quad (19)$$

This ensures that the Lyapunov drift is driven toward stability while simultaneously optimizing latency and energy consumption. Since the severity-aware virtual queue is updated using bounded QVSI values, repeated minimization of the drift guarantees that the time-average queue backlog remains finite, as expressed in Eq. (20).

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[Z_i(t)] < \infty \quad (20)$$

which implies mean-rate stability of all virtual queues. Thus, the proposed framework ensures long-term queue stability under stochastic task arrivals and time-varying wireless channels while jointly optimizing QoS violation severity, latency, and energy consumption in real time.

6) Severity-aware offloading decision policy: Based on the instantaneous system state, each IoT task chooses one of the execution modes among local processing, MEC offloading, or cooperative MEC execution. The decision is formulated as an LDPP minimization problem that combines queue stability, latency and energy consumption.

Let the three binary decision variables $x_i^{loc}(t)$, $x_i^{mec}(t)$, and $x_i^{coop}(t)$ for the three modes of execution. The constraint is enforced by the system, which is represented by Eq. (21).

$$x_i^{loc}(t) + x_i^{mec}(t) + x_i^{coop}(t) = 1, \forall_i \in N \quad (21)$$

At each time slot t, the system solves the following instantaneous optimization problem under the LDPP framework as shown in Eq. (22).

$$\min_{x_i^{loc}(t), x_i^{mec}(t), x_i^{coop}(t)} \Delta L(t) + V \cdot \mathbb{E}[\alpha E(t) + \beta T(t) \mid Z(t)] \quad (22)$$

subject to execution mode constraint (15), computation resource constraint: $\sum f_i(t) \leq F^{max}$, communication bandwidth constraint: $\sum b_i(t) \leq B^{max}$, binary decision constraints: $x_i^{loc}(t), x_i^{mec}(t), x_i^{coop}(t) \in \{0,1\}$. Where $Z(t)$ represents the severity-aware virtual queue vector. Tasks with higher queue backlog values $Z_i(t)$ are implicitly prioritized due to their stronger contribution to the Lyapunov drift term. The optimal decision is obtained by evaluating the drift-plus-penalty cost corresponding to each execution mode and selecting the mode that minimizes the objective function. This converts the long-term stochastic optimization problem into a real-time per-slot control policy suitable for dynamic 5G/B5G MEC environments.

In local execution, tasks are processed directly at the IoT device using its available computational resources. In MEC offloading, tasks are transmitted to the nearest edge server via the base station for remote execution. In cooperative MEC execution, tasks are partitioned into smaller subtasks and distributed across multiple MEC servers according to their current computational load and queue states. Coordination among MEC servers is achieved through lightweight exchange of control information, such as queue length and resource availability, via an edge orchestration layer. This coordination introduces negligible overhead compared to task execution time and is not explicitly modeled in delay analysis. Inter-edge communication is assumed to occur over a high-speed wired backhaul, and its delay is incorporated into the overall MEC execution delay while remaining significantly lower than wireless transmission delay. The proposed approach dynamically responds to channel conditions, queue states, and task urgency, allowing real-time severity-aware offloading, while reducing the severity of QoS violation, energy consumption, and latency.

7) Task execution and system operation: After the offloading decision $x_i^{loc}(t)$, $x_i^{mec}(t)$, and $x_i^{coop}(t)$ is determined, each task on the IoT is implemented based on the selected mode. In the local execution mode, the tasks are executed locally at the IoT device using its available computational resources. In the MEC offloading mode, tasks are transmitted to the nearest edge server via the base station, while in the cooperative MEC mode, tasks are distributed among multiple edge servers for load balancing. The overall delay of the executions is controlled by a transmission delay and a computation delay, depending on the mode used. Each task requires a certain amount of energy given by Eq. (23).

$$E_i(t) = \begin{cases} E_i^{loc}(t), & if\ executed\ locally \\ E_i^{tx}(t) + E_i^{mec}(t) & if\ offloaded\ to\ MEC \end{cases} \quad (23)$$

where, $E_i^{tx}(t)$ is the transmission energy, $E_i^{loc}(t)$ is the local computation energy, and $E_i^{mec}(t)$ is the energy consumption at MEC. For cooperative MEC execution, the total energy consumption is modeled as the sum of distributed computation energy across participating MEC servers, along with the corresponding (aggregated) communication cost among them. Lastly, the proposed severity model is used to evaluate the severity of QoS violation $S_i(t)$ to represent performance degradation. All metrics are stored over time and evaluated the system performance under the dynamic 5G/B5G IoT environments.

8) Queue update and continuous operation: After task execution, the severity-aware virtual queues are updated to reflect the new observed level of QoS violations, which increases the contribution of tasks with higher delays or recurring violations to the future scheduling decisions. The system operates in a continuous time-slot-based manner where the arrival of IoT tasks, channel conditions, and resources change dynamically over time. At every time slot, updates value of the queues and system parameters are returned to the Lyapunov-based optimization framework, allowing adaptive

and real-time offloading decisions. This continuous feedback mechanism ensures the proposed framework remains responsive to the changing network conditions in 5G/B5G IoT environments while maintaining the long-term stability of the system. As a result, the system effectively balances the tradeoff between latency, energy consumption, and severity-aware QoS requirements, ensuring high performance in highly dynamic and heterogeneous IoT workloads.

IV. Experimental Analysis and Discussion

The simulation is implemented in a Python-based discrete-time environment that models a 5G/B5G IoT network consisting of 100 heterogeneous IoT devices and 5 MEC servers over 1000 time slots. Each IoT device generates stochastic computation tasks with varying task sizes, CPU cycle requirements, latency deadlines, and criticality levels. The task arrivals are modeled using a Poisson arrival process with an arrival rate ranging from 0.05 to 0.15 tasks/ms, which captures the stochastic nature of heterogeneous IoT traffic. Task sizes are uniformly distributed between 0.1 MB and 10 MB, while CPU cycle requirements and latency deadlines are randomly assigned within the ranges listed in Table I to represent diverse URLLC and mMTC applications. The simulation primarily considers independent stochastic task arrivals rather than bursty or temporally correlated traffic, enabling a controlled evaluation of the proposed severity-aware MEC offloading framework under heterogeneous workload conditions. To ensure statistical reliability, all results are obtained by performing multiple independent simulation runs under identical settings, and the final reported values are computed as the average over repeated trials with different random seeds.

TABLE I. System Hyperparameters used in the Proposed Severity-Aware MEC Offloading in 5G/B5G IoT Environment

Category	Parameter	Value
Lyapunov Control Parameters	V (drift penalty weight)	10
	α (energy weight)	0.5
	β (latency weight)	0.5
	Queue update factor	severity-based (QVSI)
QVSI (Severity Model)	Criticality weights	High=3, Medium=2, Low=1
Network / Communication Settings	Bandwidth	20 MHz
	Transmission rate	100 Mbps
	Tx power	0.5 W
	Channel model	stable AWGN (typical assumption)
Compute Resources	IoT Device CPU	1 GHz
	MEC Server CPU	20 GHz
	MEC servers	5 nodes
	Task arrival rate	0.05 – 0.15 tasks/ms
Task Model	Task size	0.1 – 10 MB
	CPU cycles	0.5 – 2 GHz cycles equivalent
	Deadline range	50 – 500 ms
	Task types	URLLC + mMTC mixed

This approach captures system variability arising from stochastic task arrivals, channel conditions, and queue dynamics, and improves the robustness of the performance evaluation. For each task, both local execution delay and MEC execution delay, including transmission, computation, and queueing delays, are estimated to emulate realistic MEC operations. Table I details the parameters employed in the experimental analysis.

A. Performance Evaluation of the Proposed Framework

The performance evaluation reveals that the proposed MEC offloading mechanism with severity awareness results in substantial improvements in terms of the various QoS metrics and the efficiency of the system usage scenarios for the 5G/B5G IoT environment. The system has an average task delay of 82ms and a maximum worst-case delay of 280ms. This indicates that the latency can be controlled even during high network traffic.

The 4.8% QOS violation rate confirms the effectiveness of the QoS Violation Severity Index (QVSI) to reduce deadline misses. The 98.5% SLA satisfaction rate shows reliable service delivery for different IoT applications. The obtained system throughput of 103 tasks per second shows strong scalability when tasks are densely injected into the system. Energy efficiency is also significantly improved, with an average consumption of 6.0mJ per slot due to optimized Lyapunov drift-plus-penalty control. The workload distribution of the CPU between MEC (70%) and IoT devices (38%) shows effective workload offloading and a lower burden on the IoT devices.

The Jain fairness index value of 0.955 indicates that the resources are distributed fairly for the tasks. An offloading ratio of 72% shows that most of the tasks get completed through the utilization of MEC. The queue lengths are observed as stable, being between 2.8 and 5.0, which suggests stability and convergence in the network. This proves that the proposed framework maintains controlled congestion and efficient long-term operation.

1) Overall system performance evaluation:

Fig. 4 evaluates the proposed severity-aware Lyapunov-based MEC offloading algorithm in 5G/B5G IoT networks.

Fig. 4(a) shows that the system throughput steadily increases. This proves the efficiency of the framework to manage tasks and improve resource efficiency.

Fig. 4(b) exhibits a steady decline in average energy consumption per slot. This indicates the energy-efficient offloading decisions.

Fig. 4(c) illustrates a decrease in QoS violation rate. It validates the effectiveness of the proposed QVSI and severity-aware optimization.

Fig. 4(d) represents a rapid rise in the task completion rate. It proves improved reliability and latency control.

Fig. 4(e) shows a very high Jain's Fairness Index, which confirms fairness in resource sharing. The results prove that the proposed framework successfully minimizes severe QoS violations while improving throughput, energy efficiency, reliability, and fairness in MEC-enabled 5G/B5G IoT systems.

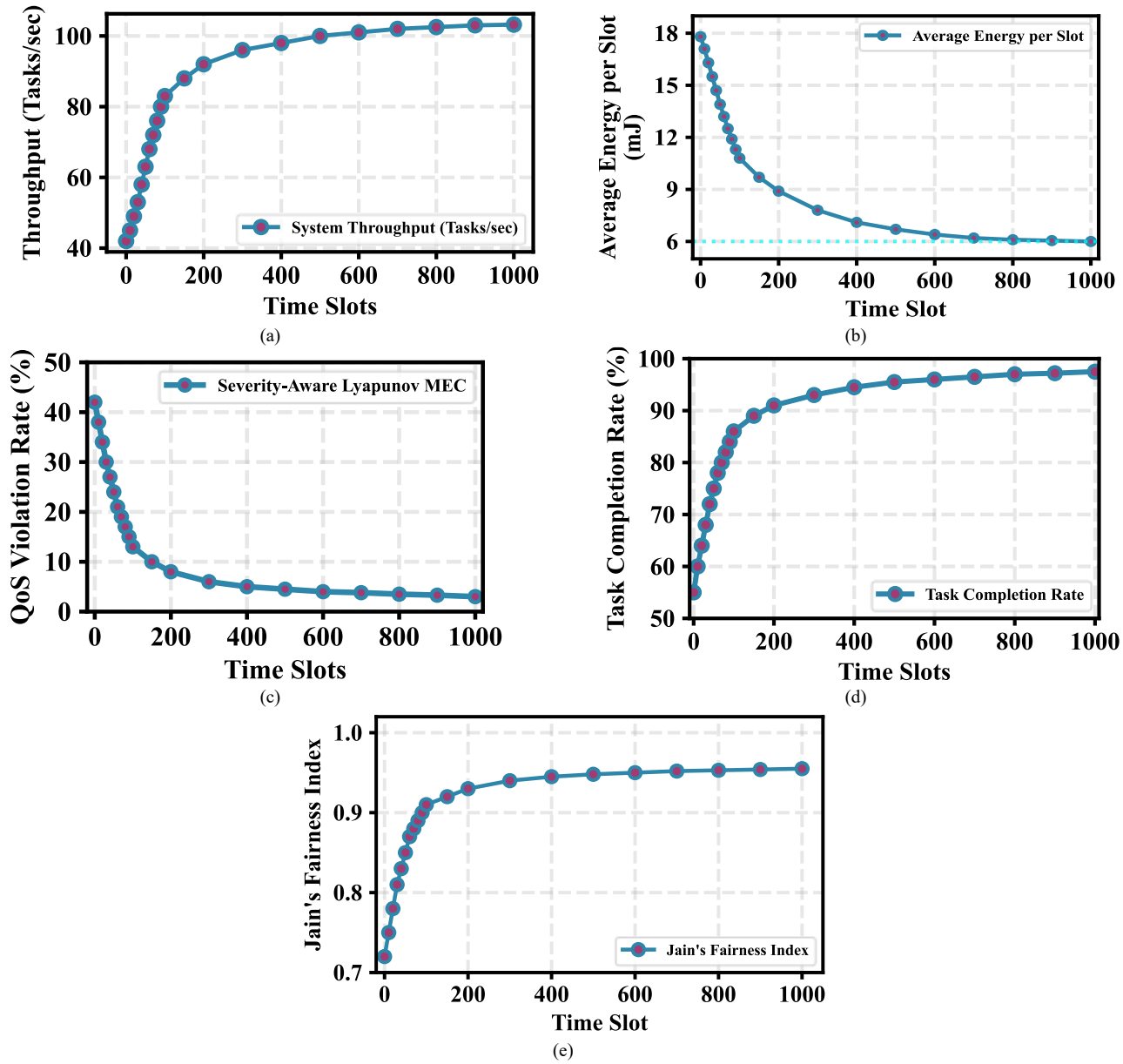


Fig. 4. Performance evaluation of the proposed severity-aware Lyapunov MEC offloading framework: (a) system throughput, (b) average energy per slot, (c)QoS violation rate, (d) task completion rate, and (e) Jain's fairness index over time slots.

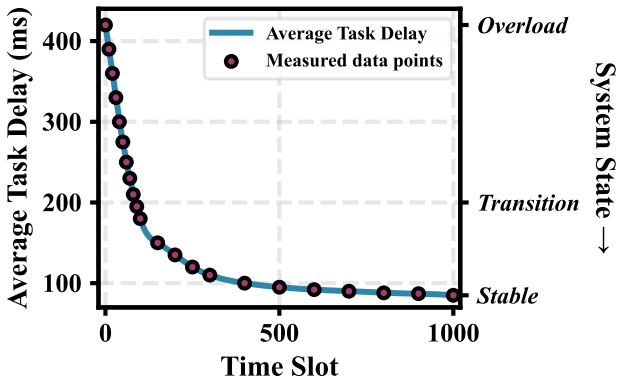


Fig. 5. Convergence behavior of the proposed severity-aware Lyapunov MEC offloading algorithm.

2) Convergence and queue stability analysis: Fig. 5 shows how effectively the proposed severity-aware LDPP MEC offloading approach converges in a 5G/B5G IoT environment. Initially, the system operates under overload conditions with a high average task delay. As the optimization process progresses, the severity-aware virtual queues stabilize. Hence, the delay is significantly reduced as we move towards the stability period. Finally, the system converges to a steady state with minimal delay. This proves that the proposed approach provides minimal task delays, stability, and effective MEC offloading performance for heterogeneous IoT applications.

Fig. 6 depicts the normalized virtual queue length change under the severity-aware Lyapunov MEC offloading framework in a 5G/B5G IoT environment. When overload occurs, the queue

rapidly increases because of the congestion and high task arrival rate in the overload phase. As the Lyapunov Drift-Plus-Penalty optimization becomes active, the queue length gradually decreases and converges toward a stable low-backlog state.

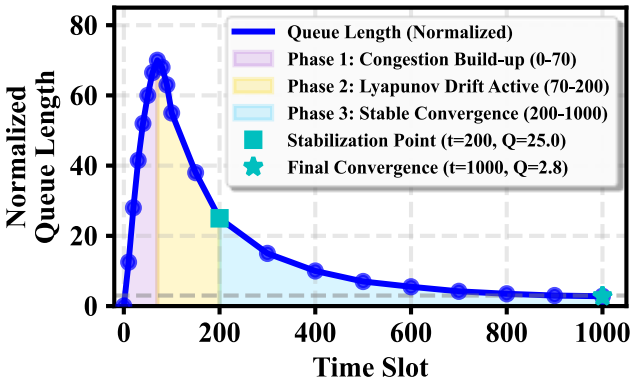


Fig. 6. Convergence of severity-aware virtual queue length in the proposed Lyapunov MEC framework.

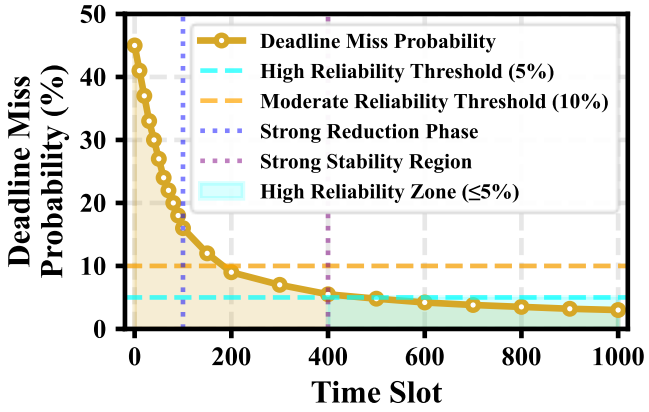


Fig. 7. Temporal reduction of deadline miss probability in the proposed severity-aware MEC offloading framework.

The stabilization and final convergence points indicate good queue stabilization and system stability. This validates the proposed algorithm's ability to overcome congestion, to achieve congestion-aware queue stabilization, and ensure reliable MEC task scheduling under dynamic IoT workloads.

Fig. 7 depicts the assessment of the proposed severity-aware Lyapunov-based MEC offloading method using the evaluation of deadline miss probability over 1000 time slots in a 5G/B5G IoT environment.

The system starts with relatively large deadline miss probabilities due to large task arrivals and queue instabilities. However, as the optimization framework adapts, the deadline misses probability falls sharply in the strong reduction period. It provides efficient decisions for task scheduling and offloading. The system reaches the optimal high-reliability state after around 400 time slots, with the probability of missing deadlines kept below 5%. This shows that the proposed QVSI-based approach reduces severe QoS violations. The results assure a high reliability and stability level for heterogeneous and latency-sensitive IoT systems. The performance of the proposed severity-aware Lyapunov-based MEC offloading approach in mitigating the severity of QoS violation has been illustrated in Fig. 8. The average value of QVSI is large in the beginning because of heavy deadline violations.

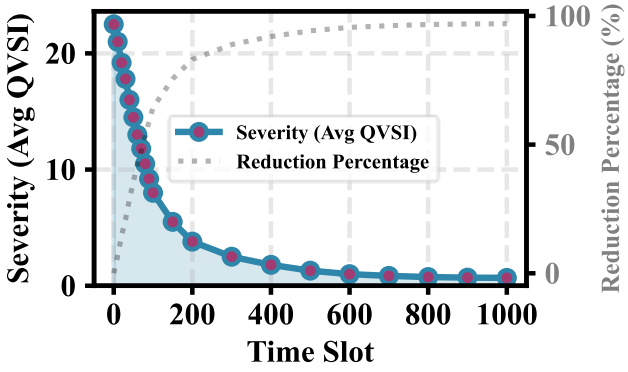


Fig. 8. Convergence of average QoS violation severity (QVSI) and reduction percentage over 1000 time slots.

3) QoS violation severity and reliability analysis: But it rapidly decreases as the algorithm learns optimal offloading decisions and prioritizes critical IoT tasks. Simultaneously, the reduction percentage steadily increases from approaching near-optimal performance. This proves the proposed framework reduces severe QoS violations, stabilizes the system, and improves service reliability for heterogeneous latency-sensitive IoT applications in 5G/B5G MEC networks while maintaining resource management. The distribution of IoT tasks is shown in Fig. 9, according to different QVSI ranges in the proposed severity-aware MEC framework.

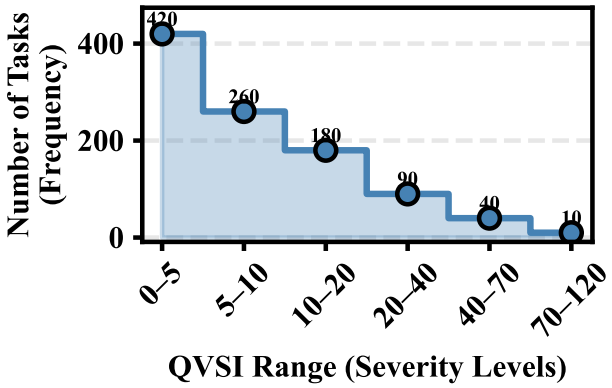


Fig. 9. Frequency distribution of QoS violation severity index (QVSI) across IoT tasks in the proposed severity-aware Lyapunov-based MEC offloading framework.

The majority of tasks are in the low-severity range (0-5), which indicates that most tasks do not suffer from high or even any deadline violations. The number of tasks decreases as the QVSI level increases, with only a few tasks suffering from severe QoS violations above 70. This shows the capability of the proposed Lyapunov-based MEC offloading algorithm in mitigating critical cases of deadline misses and prioritizing critical IoT tasks in congested 5G/B5G network environments.

4) Delay performance analysis under dynamic network conditions: Fig. 10 shows the average task delay for executing

the task locally and offloading it to the MEC in 5G/B5G IoT networks for different network loads. As the level of network load increases from low load to maximum congestion, the delay of local execution rises from 380ms to 620ms due to limited resource capabilities. In comparison, MEC offloading has lower delays, which grow incrementally from 120ms to 210ms. This shows that MEC-enabled offloading effectively reduces latency and improves the QoS performance under heavy traffic conditions, thus supporting latency-sensitive and critical IoT applications.

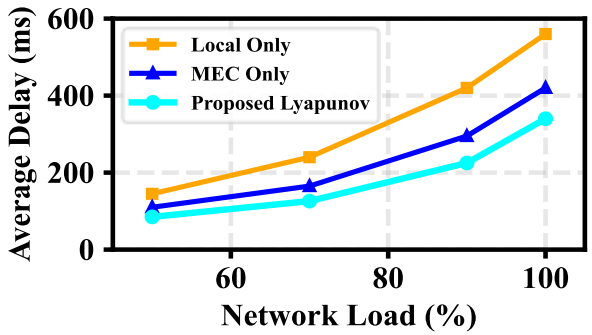


Fig. 10. Performance comparison of local computing, MEC offloading, and proposed severity-aware Lyapunov MEC offloading algorithm in terms of average delay under increasing network load.

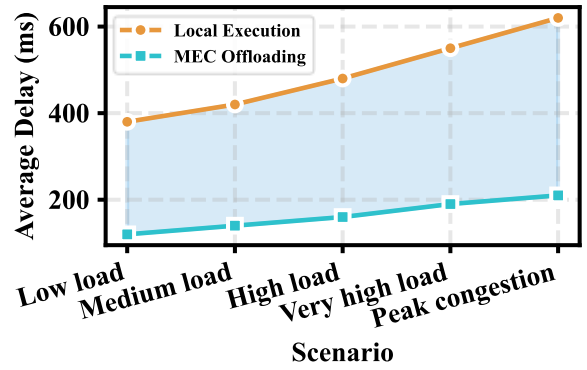


Fig. 11. Average delay comparison between local execution and MEC offloading under different network load conditions.

Fig. 11 shows the average task delay versus the increasing network load for Local Only, MEC Only, and the proposed Lyapunov scheme. The delay of all schemes increases with the network load due to congestion and limited resources. The Local Only scheme has the largest delay due to the low processing capability of IoT devices. The MEC-only scheme reduces the delay using the edge servers but still suffers from the heavy load. The proposed severity-aware Lyapunov algorithm achieves the lowest delay by intelligently balancing local execution and MEC offloading while prioritizing critical tasks. This proves that the proposed framework provides better QoS and latency performance in 5G/B5G IoT networks.

Fig. 12 depicts the delay variation of critical and non-critical IoT tasks over different time slots with the proposed framework. At first, the critical tasks suffer from higher delay due to their strict QoS requirement and higher processing priority, but as time progresses, the delays for both types of tasks are significantly reduced due to the severity-aware Lyapunov optimization that efficiently handles the queue congestion and resource allocation. The delay gap between critical and non-critical tasks also reduces over time, which depicts fair and stable scheduling.

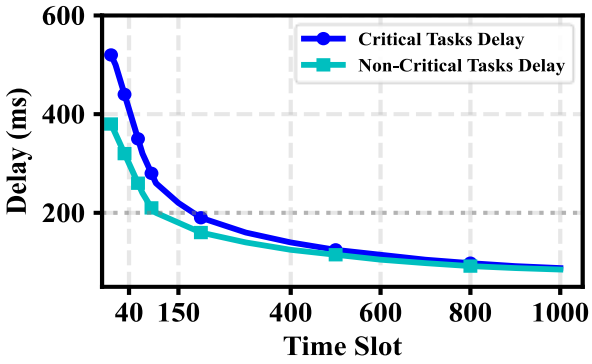


Fig. 12. Delay performance comparison between critical and non-critical IoT tasks over time slots using the proposed severity-aware Lyapunov MEC offloading framework in 5G/B5G networks.

This validates the effectiveness of the proposed algorithm to prioritize critical tasks while simultaneously maintaining the overall QoS stability and low latency in 5G/B5G IoT networks.

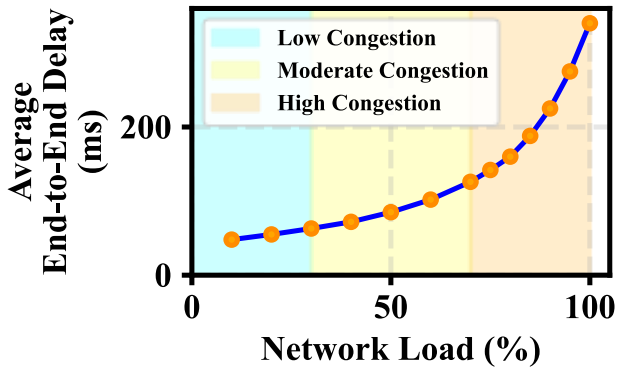


Fig. 13. Impact of network congestion levels on average end-to-end delay in the proposed severity-aware MEC offloading framework for 5G/B5G IoT networks.

Fig. 13 shows how the average end-to-end delay increases with network load under low, moderate, and high congestion conditions. In low congestion, the delays will be kept to a minimum due to abundant network and MEC resources. With an increase in congestion in terms of low, medium, and high, the delay will increase because of the increased queue formation, transmission delays, and scarcity of resources. The sudden jump towards the point where the network is full denotes the impact of heavy congestion on the QoS. This proves network congestion affects delay and demonstrates the importance of the proposed severity-aware Lyapunov MEC offloading framework in managing congestion and maintaining QoS in 5G/B5G IoT systems.

5) Energy efficiency analysis: Fig. 14 compares the comparison between the energy consumption for computation and transmission of the Local offloading, the conventional MEC, and the newly proposed severity aware MEC offloading

strategy. The results show that local execution consumes the highest computation energy (18.5mJ), while the proposed method reduces both computation and transmission energy compared to conventional MEC. This proves that the modified Lyapunov-based severity-aware offloading algorithm efficiently balances local processing and MEC offloading decisions, minimizing overall energy consumption while maintaining QoS requirements for heterogeneous 5G/B5G IoT applications.

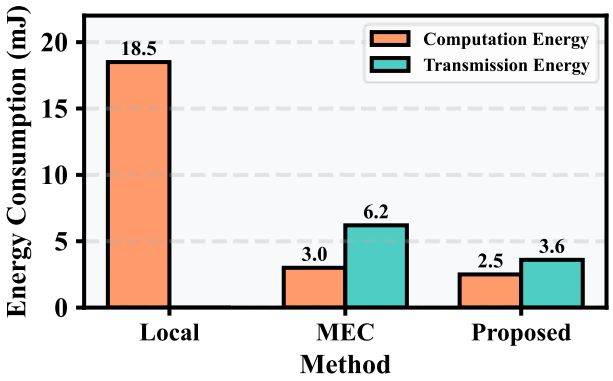


Fig. 14. Energy consumption comparison of local execution, conventional MEC, and proposed severity-aware MEC offloading method.

6) Sensitivity analysis: Table II presents a sensitivity analysis of the Lyapunov control parameter V on key performance metrics, including average task delay, energy consumption, QoS violation rate, and Jain's fairness index. As V increases from 1 to 50, the system achieves improved energy efficiency and reduced QoS violation rates due to stronger emphasis on penalty optimization in the Drift-Plus-Penalty framework.

TABLE II. SENSITIVITY ANALYSIS OF LYAPUNOV CONTROL PARAMETER V

V Value	Average Task Delay (ms)	Energy Consumption (mJ)	QoS Violation Rate (%)	Jain's Fairness Index
1	96	6.8	7.1	0.931
5	88	6.3	5.6	0.944
10	82	6.0	4.8	0.955
20	80	5.9	4.6	0.952
50	79	5.8	4.5	0.947

However, higher values of V slightly affect delay performance and fairness after a threshold, indicating a trade-off between latency minimization and system stability. The results demonstrate that $V = 10\text{-}20$ provides a balanced performance across all metrics.

B. Ablation Analysis

Table III evaluates the contribution of each major component of the proposed severity-aware MEC offloading framework. Different models are compared with the complete MEC offloading framework. The removal of the QVSI variant shows higher delay, energy consumption, and QoS violation rate, which proves the significance of the QVSI for quantifying and minimizing severe deadline violations. Likewise, removing virtual queues based on severity causes an improvement in throughput and fairness, highlighting that severity queue modeling improves the stability of a virtual queue and prioritizes critical tasks.

TABLE III. ABLATION ANALYSIS OF THE PROPOSED SEVERITY-AWARE MEC OFFLOADING FRAMEWORK IN 5G/B5G IOT NETWORKS

Method Variant	Average Task Delay (ms)	System Throughput (tasks/sec)	Average Energy Consumption (mJ)	Jain's Fairness Index	QoS Violation Rate (%)
w/o QVSI	105.0 ± 5.8	88.0 ± 3.2	7.8 ± 0.4	0.910 ± 0.010	9.5 ± 0.7
w/o Severity Queues	112.0 ± 6.1	85.0 ± 3.0	8.2 ± 0.5	0.895 ± 0.012	11.2 ± 0.8
w/o LDPP	128.0 ± 7.0	79.0 ± 3.5	9.6 ± 0.6	0.860 ± 0.015	14.8 ± 1.0
w/o Task Criticality	118.0 ± 6.5	83.0 ± 3.1	8.7 ± 0.5	0.882 ± 0.011	12.6 ± 0.9
Local Execution Only	155.0 ± 8.2	65.0 ± 2.8	12.5 ± 0.8	0.790 ± 0.018	22.0 ± 1.3
Full Proposed Model	**82.0 ± 4.6**	**103.0 ± 3.4**	**6.0 ± 0.3**	**0.955 ± 0.008**	**4.8 ± 0.5**

By removing the LDPP configuration, the worst performance is obtained among the algorithmic variants in terms of delay and QoS violation rate, showing that the optimal configuration of LDPP is a crucial and effective component for efficient online offloading and resource management. Without the Task Criticality model, performance begins to drop significantly, especially in fairness and QoS violation rate, indicating the significance of addressing heterogeneous task priority in latency-sensitive IoT applications. The Local Execution Only baseline also suffers from the highest delay, energy consumption, and QoS violations as tasks are executed only on IoT devices without the support of MEC, causing resource congestion and inefficient task execution. In contrast, the full proposed model achieves the highest energy efficiency, the highest throughput, the lowest delay, improved fairness, and a minimal QoS violation rate. These results show that the combination of using QVSI, severity-aware queues, Lyapunov optimization, and task criticality improves the reliability and QoS performance in 5G/B5G MEC-supported IoT networks.

C. Comparative Analysis

To ensure a fair comparison, all existing methods are implemented under identical simulation settings as the proposed framework. Table IV shows a comparison between the proposed severity-aware MEC offloading framework and existing state-of-the-art methods in terms of task delay, throughput, energy consumption, fairness, and QoS violation rate in 5G/B5G IoT networks. The results indicate that the traditional methods, such

as DRL-DDQN, HATO, PSO, Lyapunov optimization with PMP, and DNNs-RL, result in moderate performance gains, but still incur high latency and QoS violations due to limited accounting for the severity of the violation and the criticality of the task. Conversely, the proposed method achieves the lowest average task delay (82ms), minimum energy consumption (6.0mJ), the highest throughput (103 tasks/sec), and the best fairness index (0.955), and also reduces the rate of violation of QoS to 4.8%. The improvements demonstrate that the proposed framework enables more efficient and reliable MEC task offloading in 5G/B5G environments for heterogeneous and latency-sensitive IoT applications.

TABLE IV. COMPARATIVE PERFORMANCE ANALYSIS OF THE PROPOSED SEVERITY-AWARE MEC OFFLOADING METHOD AGAINST EXISTING APPROACHES

Reference	Method	Average Task Delay (ms)	System Throughput (tasks/sec)	Average Energy Consumption (mJ)	Jain's Fairness Index	QoS Violation Rate (%)
Zhai et al. [25] (2024)	DRL-DDQN	95.0 ± 5.2	92.0 ± 3.0	7.2 ± 0.4	0.920 ± 0.009	8.5 ± 0.6
Benbraika et al. [26] (2024)	HATO	110.0 ± 6.0	85.0 ± 2.9	8.5 ± 0.5	0.900 ± 0.011	11.0 ± 0.8
Alam et al. [27] (2024)	PSO	108.0 ± 5.8	87.0 ± 3.1	8.1 ± 0.4	0.905 ± 0.010	10.2 ± 0.7
Luo et al. [29] (2025)	Lyapunov optimization with PMP algorithm	100.0 ± 5.5	90.0 ± 3.0	7.6 ± 0.4	0.915 ± 0.009	9.0 ± 0.6
Yang et al. [30] (2024)	DNNs-RL	92.0 ± 5.0	94.0 ± 3.2	7.0 ± 0.3	0.925 ± 0.008	7.8 ± 0.5
Proposed	**LDPP**	**82.0 ± 4.6**	**103.0 ± 3.4**	**6.0 ± 0.3**	**0.955 ± 0.008**	**4.8 ± 0.5**

D. Discussion

The complex environment of 5G and B5G IoT introduces significant challenges in designing an efficient and reliable MEC task offloading strategy. Current frameworks mainly use binary models of QoS evaluation, which do not consider task performance as being either satisfied or violated. This approach fails to capture the level of deadline violations, especially for latency-sensitive and mission-critical IoT applications. Furthermore, most of the conventional approaches consider optimizing mean QoS metrics without the explicit consideration of extreme QoS degradation scenarios. Moreover, lack of information about task heterogeneity, in particular the criticality of the tasks, often leads to unfair scheduling decisions and poor performance of high-priority tasks. The dynamic nature of networks, such as varying wireless channels, computation loads, and resource-limited IoT devices, further complicates efficient task offloading and requires relying on static decision-making strategies in real-world MEC systems. To tackle these challenges, this study introduces a severity-aware MEC task offloading framework for heterogeneous 5G/B5G IoT environments. The main idea is to introduce a QVSI that not only estimates whether a task misses its deadline but also the magnitude of the violation, while including task criticality in the evaluation process. This achieves a more realistic representation of QoS degradation in the real IoT scenario. Moreover, a severity-aware virtual queue model is designed to capture system backlog as violation intensity instead of just tasks, enabling fine-grained control of system stability. These components are combined in a modified LDPP optimization framework to allow real-time and adaptive offloading decisions. The framework jointly minimizes latency, energy, and QoS violation severity while ensuring long-term stability of queues under dynamic network conditions.

The proposed objectives are verified by extensive simulation experiments conducted in a Python-based discrete-time MEC environment. The results show that the framework has significant improvements in system performance across various metrics. The average task delay is decreased to 82ms, and the mean energy consumption is minimized to 6.0mJ/slot. The deadline miss rate is significantly reduced to 4.8%, indicating an effective deadline miss mitigation. Moreover, the system achieves an efficient and fair performance with a Jain's fairness index of 0.955 and a throughput of 103 tasks/second among the heterogeneous IoT devices. The QoS violation severity is reduced by about 97%, confirming good convergence performance and the effectiveness of the severity-aware Lyapunov optimization mechanism in reducing the severity of the QoS violation under various workloads.

Overall, these results validate that the proposed model achieves its objectives of minimizing severe QoS violations while optimizing latency, energy efficiency, and fairness. However, the proposed framework relies on accurate knowledge of system state information, such as channel conditions, task characteristics, and computational requirements, which may not always be fully reliable in highly dynamic or partially observable environments. In addition, the current study assumes an AWGN-based communication model for performance evaluation, which provides a simplified and controlled baseline but does not capture realistic wireless fading dynamics in practical 5G/B5G systems. Future work could include the application of learning techniques such as deep reinforcement learning (DRL) or federated learning to improve adaptability under uncertainty and incomplete system information. Moreover, the framework can be extended to more realistic wireless channel models such as Rayleigh and Rician fading, as well as real-world edge computing testbeds. Mobility-aware and energy harvesting IoT models can also be incorporated to further enhance the applicability, robustness, and practical utility of the proposed framework for future 5G/B5G networks.

V. CONCLUSION

This work proposes a severity-aware MEC task offloading framework for 5G/B5G IoT that effectively resolves the limitations of the conventional binary QoS and non-criticality-aware task offloading methods. The framework introduces a QVSI and severity-aware virtual queuing integrated with a modified LDPP optimization model for real-time adaptive offloading decisions. This allows joint optimization of latency,

energy, and violation of QoS, and ensures stability of the system under dynamic network conditions. The excellent performance of the proposed approach is confirmed through simulation, where the average task delay is 82 ms, the energy consumption per slot is 6.0 mJ, the violation of the QoS of 4.8%, the system throughput is 103 tasks/sec, and the fairness index is 0.955. The significant decrease in QoS severity of violation illustrates strong convergence and reliability. Overall, the proposed framework ensures efficient, stable, and severity-aware resource allocation in MEC-enabled 5G/B5G IoT, offering a promising solution for future intelligent edge computing applications.

STATEMENTS AND DECLARATIONS

Author contributions: All authors contributed to the conception of the problem setting and overall design of the work. S.S.R built the conceptualization and methodology. R.S. implemented the work and writing. Authors involved in the visualization of concepts. This version was approved by all authors, who also read and approved the manuscript.

Funding: No funding was received for conducting this study.

Conflict of interest: The authors declare that they have no conflict of interest.

Availability of data and materials: The data set is not available; only a simulation with parameters is used.

Ethical approval: The research is original, and the authors of this manuscript created all the figures and tables.

Consent to participate: Not applicable.

Consent for publication: All authors agree with the submission of the manuscript to this journal and possible publication afterwards.

REFERENCES

[1] M. H. Alanazi, “Machine Learning-based Secure 5G Network Slicing: A Systematic Literature Review,” Int. J. Adv. Comput. Sci. Appl., vol. 14, pp.377, 2023.

[2] C. Zhiyi, M. Aman, and A. Hafizah, “A Review on NS Beyond 5G: Techniques, Applications, Challenges and Future Research Directions,” Int. J. Adv. Comput. Sci. Appl., vol. 15, pp. 117, 2024.

[3] T. Senevirathna, V. H. La, S. Marcha, B. Siniarski, M. Liyanage, and S. Wang, “A survey on XAI for 5G and beyond security: Technical aspects, challenges and research directions,” IEEE Commun. Surv. Tutor., vol. 27, pp. 941–973, Aug. 2024.

[4] Kaur and A. Godara, “Machine learning empowered green task offloading for mobile edge computing in 5G networks,” IEEE Trans. Netw. Serv. Manag., vol. 21, pp. 810–820, July 2023.

[5] L. Peng, P. H. Ho, and K. Zhao, “Task offloading in terrestrial-support-free multi-layer multi-access edge computing,” IEEE Commun. Mag., vol. 62, pp. 82–87, July 2024.

[6] M. A. Khan, N. Kumar, S. H. Alsamhi, G. Barb, J. Zywiołek, I. Ullah, F. Noor, J. A. Shah, and A. M. Almuhaideb, “Security and privacy issues and solutions for UAVs in B5G networks: A review,” IEEE Trans. Netw. Serv. Manag., vol. 22, pp. 892–912, Oct. 2024.

[7] A. Darwin Jose Raju, S. Solai Manohar, T. C. Jermin Jeaunita, Energy Prediction and Scheduling in Battery-Less Backscatter Sensor Network for Sustainable Network Management, Pervasive and Mobile Computing, 2026, 102234, ISSN 1574-1192.

[8] R. Huang, W. Wen, X. Chen, Z. Zhou, Q. Chen, and C. Dong, “Joint power allocation and task replication for reliability-sensitive services in NOMA-enabled vehicular edge computing,” IEEE Trans. Veh. Technol., vol. 73, pp. 4178–4193, Oct. 2023.

[9] H. Guo, Y. Wang, J. Liu, and C. Liu, “Multi-UAV cooperative task offloading and resource allocation in 5G advanced and beyond,” IEEE Trans. Wirel. Commun., vol. 23, pp. 347–359, May 2023.

[10] S. Awoyemi and B. T. Maharaj, “Adaptive power management for multiaccess edge computing-based 6G-inspired massive Internet of Things,” IET Wirel. Sens. Syst., vol. 15, p. e70000, Jan. 2025.

[11] D. Logeshwari, S. Asha, T. C. Jermin Jeaunita, M. Povaneswari, "Proactive Channel Assignment and Modified CR-AODV for Stability-Driven Multi-Hop Routing in Cognitive Radio Networks," Radioengineering, vol. 34, Dec 2025.

[12] E. F. Maleki, W. Ma, L. Mashayekhy, and H. J. La Roche, “QoS-aware content delivery in 5G-enabled edge computing: Learning-based approaches,” IEEE Trans. Mob. Comput., vol. 23, pp. 9324–9336, Feb. 2024.

[13] H. W. Kao and E. H. K. Wu, “QoE sustainability on 5G and beyond 5G networks,” IEEE Wirel. Commun., vol. 30, pp. 118–125, Mar. 2023.

[14] G. Nieto, I. De la Iglesia, U. Lopez-Novoa, and C. Perfecto, “Deep reinforcement learning techniques for dynamic task offloading in the 5G edge-cloud continuum,” J. Cloud Comput., vol. 13, pp. 94, May 2024.

[15] A. Elgendy, S. Meshoul, and M. Hammad, “Joint task offloading, resource allocation, and load-balancing optimization in multi-UAV-aided MEC systems,” Appl. Sci., vol. 13, pp. 2625, Feb. 2023.

[16] C. Xu, M. Lv, K. Zhang, K. Cao, G. Wang, M. Wei, and B. Peng, “Energy consumption and time-delay optimization of dependency-aware tasks offloading for industry 5.0 applications,” IEEE Trans. Consum. Electron., vol. 70, pp. 1590–1600, Dec. 2023.

[17] P. F. Moshiri, M. Simsek, and B. Kantarci, “Joint optimization of completion ratio and latency of offloaded tasks with multiple priority levels in 5G edge,” IEEE Trans. Netw. Serv. Manag., vol. 22, pp. 1357–1371, Jan. 2025.

[18] A. Abba Ari, F. Samafou, A. Ndam Njoya, A. C. Djedouboum, M. Aboubakar, and A. Mohamadou, “IoT-5G and B5G/6G resource allocation and network slicing orchestration using learning algorithms,” IET Netw., vol. 14, pp. e70002, Jan. 2025.

[19] P. Li, Y. Wang, Z. Wang, T. Wang, and J. Cheng, “Joint task offloading and resource allocation strategy for hybrid MEC-enabled LEO satellite networks: A hierarchical game approach,” IEEE Trans. Commun., vol. 73, pp. 3150–3166, Oct. 2024.

[20] N. Lassoued and N. Boujnah, “A comprehensive review of energy efficiency in 5G networks: Past strategies, present advances, and future research directions,” Computers, vol. 15, pp. 50, Oct. 2026.

[21] R. Zhang, L. Wu, S. Cao, D. Wu, and J. Li, “A vehicular task offloading method with eliminating redundant tasks in 5g HetNets,” IEEE Trans. Netw. Serv. Manag., vol. 20, pp. 456–470, Aug. 2022.

[22] A. Uddin, A. H. Sakr, and N. Zhang, “Adaptive prioritization and task offloading in vehicular edge computing through deep reinforcement learning,” IEEE Trans. Veh. Technol., vol. 74, pp. 5038–5052, Nov. 2024.

[23] F. Bahramisirat, M. A. Gregory, and S. Li, “Dynamic network slicing for resource allocation in 5G/B5G networks: An optimization-based approach,” Ad Hoc Netw., p. 104194, Feb. 2026.

[24] H. Hu, Z. Chen, F. Zhou, Z. Han, and H. Zhu, “Joint resource and trajectory optimization for heterogeneous-UAVs enabled aerial-ground cooperative computing networks,” IEEE Trans. Veh. Technol., vol. 72, pp. 8812–8826, Feb. 2023.

[25] H. Zhai, X. Zhou, H. Zhang, and D. Yuan, “Delay minimization in hybrid edge computing networks: A DDQN-based task offloading approach,” IEEE Trans. Veh. Technol., vol. 73, no. 10, pp. 15098–15108, May 2024.

[26] M. K. Benbraika, O. Kraa, Y. Himeur, K. Telli, S. Atalla, and W. Mansoor, “Enhancing 5g vehicular edge computing efficiency with the Hungarian algorithm for optimal task offloading,” Computers, vol. 13, pp. 279, Oct. 2024.

[27] A. Alam, P. Shah, R. Trestian, K. Ali, and G. Mapp, “Energy efficiency optimisation of joint computational task offloading and resource allocation using particle swarm optimisation approach in vehicular edge networks,” Sensors, vol. 24, pp. 3001, May 2024.

[28] A. Younis, S. Maheshwari, and D. Pompili, “Energy-latency computation offloading and approximate computing in mobile-edge computing

networks," IEEE Trans. Netw. Serv. Manag., vol. 21, pp. 3401–3415, Jan. 2024.

[29] K. Luo, Y. Wang, Y. Liu, and K. Zhu, "Collaborative integration of vehicle and roadside infrastructure sensor for temporal dependency-aware task offloading in the Internet of Vehicles," Int. J. Intell. Syst., vol. 2025, p. 8064086, May 2025.

[30] W. Yang, Z. Liu, X. Liu, and Y. Ma, "Deep reinforcement learning-based low-latency task offloading for mobile-edge computing networks," Appl. Soft Comput., vol. 166, p. 112164, Nov 2024.

[31] S. Mathi, G. Rohan Lal, L. C., Madala, K. A., Reddy, P. Jagadhabhiram, and G. Neelakanta Iyer, "FedBHAD: Energy-Efficient Federated Learning for Black Hole Attack Detection in RPL-Based Low-Power IoT Networks," Emerging Science Journal, vol. 9, pp.2884–2898, Dec 2025.

[32] Y. Xiao, and Y. Dong, "Optimizing AIGC Technology for IoT Devices with Deep Learning," HighTech and Innovation Journal, vol. 6, pp. 976–990, Sep 2025.

[33] H. Vidyaningtyas, H., Iskandar, A. A. Pramudita, "NOMA Performance Improvement with Downlink Sectorization," Emerging Science Journal, vol. 9, pp. 311-328, Feb 2025.

[34] A. Ateya, A. Muthanna, A. Koucheryavy, Y. Maleh, and A. A. A. El-Latif, "Energy efficient offloading scheme for MEC-based augmented reality system," Clust. Comput., vol. 26, pp. 789–806, Feb 2023.

[35] V. R. Chintapalli, R. Partani, and B. R. Tamma, "Energy efficient and delay aware deployment of parallelized service function chains in NFV-based networks," Comput. Netw., vol. 243, p. 110289, Apr. 2024.

[36] H. Yuan, T. Wang, M. Fu, and Y. Shi, "GIRP: Energy-efficient QoS-oriented microservice resource provisioning via multi-objective multi-task reinforcement learning," IEEE Trans. Mob. Comput., vol. 24, pp. 5793–5807, Mar. 2025.

[37] G. Wu, X. Chen, Y. Shen, Z. Xu, H. Zhang, S. Shen, and S. Yu, "Combining Lyapunov optimization with actor–critic networks for privacy-aware IIoT computation offloading," IEEE Internet Things J., vol. 11, pp. 17437–17452, Jan. 2024.

[38] M. An, T. Lu, X. Han, and Z. Ding, "An online distributed optimisation model solving the time-coupled conundrum by the Lyapunov drift plus penalty method," IET Gener. Transm. Distrib., vol. 19, p. e70019, Jan 2025.

Bridging Topic Modelling Outputs to Bayesian Hierarchical Model Using LLM and WordNet Parameter Labelling

Vadrianey Asas[1], Sarah Samson Juan[2]*, Stephanie Chua[3], Evan Lau[4], Jane Labadin[5]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia[1, 2, 3, 5]
Faculty of Economics and Business, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia[4]

***Abstract*—This study investigates the challenge of generating accurate and interpretable topic labels for integration into Bayesian Hierarchical Models (BHM), a critical step for interpretable probabilistic risk modelling from unstructured textual data. Using a corpus of 35,667 Malaysian business news articles published between 2019 and 2023, four topic modelling approaches, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic, were evaluated. Among these, NMF produced the most coherent and thematically consistent topics. To address the topic-labelling challenge, this study proposes an NLP-BHM framework that maps topic model outputs into a hierarchical Bayesian structure comprising interpretable topic labels and higher-level risk categories. Two semantic labelling strategies were examined: Large Language Model (LLM) prompting and WordNet-based semantic analysis. The proposed approaches enabled systematic topic interpretation and semantic clustering within the BHM framework. A case study on Malaysian business risks demonstrates that LLM-based labelling produced more coherent and contextually relevant results, while WordNet-based labelling provided a semantically consistent but vocabulary-limited alternative. Comparative results based on Mean Opinion Scores (MOS) highlight the effectiveness of LLM-based semantic labelling in improving interpretability for probabilistic business risk analysis.**

***Keywords*—Bayesian hierarchical model; natural language processing; topic modelling; large language model; WordNet**

I. Introduction

In today's data-driven economy, extracting actionable insights from unstructured text is crucial for effective risk assessment and informed decision-making. Business news articles, financial disclosures, and policy updates often contain critical early signals of systemic risks. However, their narrative form poses challenges for direct integration into quantitative models. To generate interpretable probabilistic risk representations, it is essential to translate this unstructured information into structured, interpretable inputs suitable for Bayesian Hierarchical Models (BHM).

Bayesian Hierarchical Models (BHM) provide a probabilistic framework for modelling uncertainty across multiple levels of abstraction, enabling interpretable relationships between observed variables and higher-level latent structures [1]. The power of Bayes' theorem lies in its adaptability: parameters, hypothesised relationships, and confidence intervals can all be estimated as probabilistic statements derived from the model. Traditionally, parameter labels for BHM are manually annotated, a process that can be highly time-consuming and inconsistent ([2], [3]). One notable work applies manual annotation to obtain the interpretable labels for the factor inputs in their study [4]. Despite these efforts, no significant work has integrated topic modelling outputs into BHM structures, leaving the process of bridging these outputs into BHM largely unexplored.

Prior investigations have identified the best-performing topic model based on coherence scores, which will be utilised for the integration into the BHM structure [5]. Research on topic modelling, such as the foundational work by Blei et al. [6], has established methods for extracting topics from text data, but the challenge remains in linking these outputs with BHM factor inputs. However, the core challenge in this study lies in bridging the topic outputs from the best-performing topic model with the factor inputs to obtain clear, interpretable labels required by the BHM structure.

This study aims to fill this gap by proposing a framework that integrates topic-modelling outputs into BHM structures. We evaluate two parameter labelling strategies: Large Language Model (LLM) prompting and WordNet-based semantic analysis. These strategies will help derive interpretable labels for BHM parameters, facilitating the connection between natural language processing (NLP)-derived topics and Bayesian models. The central research question guiding this work is: How can LLM-based and WordNet-based labelling strategies be leveraged to generate interpretable topic representations and construct a BHM structure for probabilistic risk analysis?

The contribution of the study lies within the framework itself. In contrast to the traditional BHM that relies on defined variables and structured inputs, the proposed approach automatically derives risk factors from unstructured business news, enabling a fully data-driven representation of emerging risk signals. In addition, the framework systematically compares two complementary semantic labelling strategies, namely LLM-based and WordNet-based approaches, to enhance the interpretability and robustness of topic-to-risk mapping. Furthermore, the model enables dynamic estimation of risk probabilities by propagating topic-level distributions across evolving news corpora, thus capturing temporal variations in business risk environments.

The study is organised as follows. In Section II, we briefly discuss related work on the approaches used in our work. We

*Corresponding author.

present our framework for integrating the topic model's output using two labelling approaches in Section III. In Section IV, we discuss the results from the Mean Opinion Scoring (MOS) evaluation and present the BHM structures built using both approaches. Finally, we present our conclusions and future work in Section V.

II. Literature Review

A. Bayesian Hierarchical Model

Bayesian modelling provides a probabilistic framework for reasoning under uncertainty by updating prior beliefs using observed evidence [2]. Bayesian Hierarchical Models (BHM) extend this framework by organising model components across multiple levels, thereby enabling information sharing and uncertainty propagation throughout the hierarchy ([2], [3]).

0 illustrates a Bayesian hierarchical structure with population-level hyperparameters, group-specific parameters, and observed data. Shared priors enable information sharing across groups, while group-level parameters capture variation and govern the observations. Posterior inference updates uncertainty across the hierarchy, supporting partial pooling and more stable estimates when group-level data are limited.

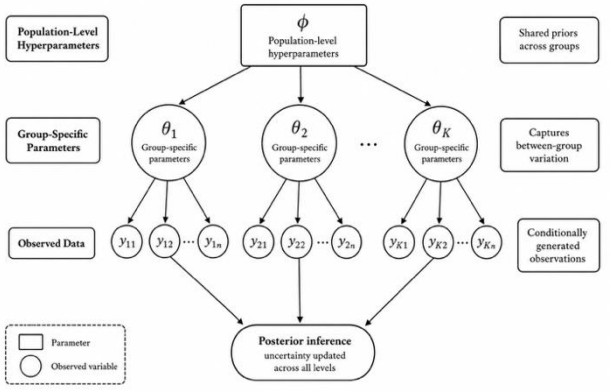


Fig. 1. A general structure of a Bayesian hierarchical model.

Various studies have applied BHM for predictive modelling across different domains. In supply chain management, BHM is used to assess disruption risks with limited data availability by estimating supplier risk probabilities and financial impacts, and incorporates downstream risk assessment measures, such as Value-at-Risk (VaR), to support decision-making [4]. Similarly, in supply chain financing, BHM improves risk evaluation by integrating macro-level credit ratings with micro-level transaction data, resulting in more accurate predictions of payment delays and revealing behavioural patterns such as correlations between delivery and payment delays [7].

Beyond financial applications, BHM has shown strong capability in ecological and environmental modelling. In species distribution studies, it integrates environmental variables with ecological interactions and latent processes, improving predictive performance when combined with complementary models such as food-web approaches [8]. In public health, BHM is applied to estimate disease burden from incomplete surveillance data, providing robust and scalable estimates for long-term monitoring and decision-making during the COVID-19 pandemic [9]. In hydrology, BHM models extreme flood risks under nonstationary conditions by integrating climate indices within a hierarchical framework, improving estimation accuracy and reducing uncertainty through partial pooling [10].

Recent studies in natural language understanding have demonstrated that uncertainty-aware probabilistic representations can better capture semantic ambiguity and subjective variation in textual interpretation, further motivating the integration of probabilistic reasoning within NLP frameworks [11].

These studies demonstrate the robustness of BHM in integrating heterogeneous data sources and modelling complex dependencies across domains. However, most existing approaches rely on predefined model structures and manually specified parameters, limiting flexibility in adapting to evolving data patterns and latent factors. This limitation reduces its effectiveness when dealing with large-scale unstructured data, where important information is often embedded within latent textual representations.

B. Topic Modelling

Topic modelling has long been a popular technique for uncovering hidden themes in large text corpora [6]. A study applies LDA to a large-scale corpus of artificial intelligence literature to extract latent topics from publication abstracts[12]. The topic distributions are then aggregated across time, journals, and countries to identify key AI subfields and analyse research trends and similarities across publication sources. Another study applies a semantically assisted NMF approach (SeNMFk) to extract latent topics and determine the optimal number of topics [13]. By integrating TF-IDF and semantic information through SPPMI matrices, the model improves topic stability and coherence compared to traditional methods.

Top2Vec has been increasingly used across domains due to its ability to combine word embeddings with clustering to produce semantically coherent topics[14]. It was applied to analyse 15,000 mobile app user feedback entries[15], where it outperformed LDA and ETM by automatically selecting the number of topics and producing more interpretable results. Evaluation metrics such as the coherence score and topic diversity demonstrated their effectiveness in capturing issues such as app glitches and performance problems. BERTopic, a transformer-based topic modelling method, has also been widely used to uncover latent themes in text data[16]. One study applied BERTopic to 2,846 PubMed abstracts on depression, anxiety, and burnout in academia[17], identifying 27 topics including pandemic-related anxiety and medical resident burnout.

A comparative study of LDA, NMF, Top2Vec, and BERTopic on COVID-19-related Twitter data highlights challenges in analysing noisy social media text and suggests that BERTopic and NMF are more effective approaches for social science applications[1]. The study also emphasises the importance of qualitative interpretation and domain expertise. However, while prior work compares topic modelling methods for extracting insights from unstructured text, most studies primarily focus on topic extraction quality, coherence, and interpretability. Limited attention has been paid to transforming latent topic representations into structured probabilistic

frameworks, such as Bayesian Hierarchical Models, for downstream risk modelling and uncertainty-aware decision-making.

C. Large Language Models

Recent advances in Large Language Models (LLMs) have significantly improved natural language understanding and semantic reasoning capabilities in NLP tasks [18]. Models such as GPT, PaLM, and LLaMA can generate context-aware responses by learning rich semantic representations from large-scale corpora [19]. Unlike traditional rule-based or ontology-driven approaches, LLMs can capture contextual relationships between words and concepts, enabling more flexible and semantically meaningful interpretation of unstructured text.

This capability has led to increasing use of LLMs in semantic extraction, topic interpretation, summarisation, and decision-support applications. Recent studies have shown that LLMs can generate coherent topic labels and semantic summaries that better align with human interpretation compared to conventional lexical resources and statistical approaches [18]. Their ability to contextualise latent textual representations makes them particularly well-suited to interpreting topic modelling outputs, where extracted keywords often lack semantic clarity when analysed in isolation. Compared to ontology-based resources such as WordNet, LLMs offer stronger contextual reasoning and semantic flexibility, especially in domain-specific environments such as finance and business analytics. Traditional lexical resources rely on predefined vocabularies and hierarchical relationships, which may limit their ability to capture emerging terminology, implicit semantics, and context-dependent meanings. In contrast, LLMs can dynamically infer semantic relationships from textual context, improving interpretability and semantic coherence in topic labelling tasks.

Despite these advantages, several challenges remain in applying LLMs within probabilistic modelling frameworks. LLM-generated outputs may vary with prompt design, sampling configurations, and model randomness, potentially affecting consistency and reproducibility. Furthermore, hallucinated or overly general outputs may reduce reliability when semantic labels are directly integrated into downstream analytical models. Consequently, human validation and structured prompting strategies remain important to ensure semantic accuracy and interpretability.

While recent work has explored the use of LLMs for semantic extraction and financial text analysis [20], limited attention has been given to integrating LLM-generated semantic representations into hierarchical probabilistic frameworks for interpretable business risk modelling. This limitation motivates the present study, which leverages LLM-based semantic labelling to bridge latent topic representations with Bayesian Hierarchical Models for uncertainty-aware risk analysis.

D. WordNet-Based Semantic Analysis

WordNet is one of the most widely used lexical databases in natural language processing for semantic analysis and knowledge representation[21]. It organises words into sets of cognitive synonyms known as synsets, which are connected through semantic relationships such as synonymy, hypernymy, and hyponymy. These structured semantic relationships enable WordNet to support tasks including word sense disambiguation, semantic similarity measurement, text classification, and ontology-based information retrieval.

In topic modelling and semantic analysis, WordNet has been widely used to improve topic coherence and interpretability by identifying semantically related concepts and hierarchical relationships between keywords. Several studies employ WordNet-based semantic similarity measures, such as Wu–Palmer similarity and path-based similarity, to cluster semantically related terms and generate interpretable topic labels [22]. By leveraging lexical hierarchies, WordNet enables semantically related topics to be grouped into broader conceptual categories, supporting more structured knowledge representation. Compared to statistical topic modelling methods, ontology-based approaches offer greater consistency and reproducibility because semantic relationships are derived from predefined lexical structures rather than probabilistic generation. This deterministic behaviour makes WordNet particularly useful in applications requiring stable semantic mappings and explainable semantic relationships. Furthermore, hierarchical lexical structures align naturally with probabilistic hierarchical models, making WordNet-based semantic grouping well-suited to constructing interpretable latent-factor hierarchies.

However, WordNet also presents several limitations, particularly in domain-specific applications such as business and financial risk analysis. Since WordNet relies on manually curated vocabularies, it may fail to capture emerging terminology, contextual nuances, and specialised business concepts commonly found in dynamic textual corpora. In addition, lexical similarity does not necessarily reflect contextual semantic meaning, which may lead to overly generic or semantically weak topic labels when applied to complex business news data.

These limitations motivate the exploration of context-aware semantic approaches that use Large Language Models to dynamically infer semantic relationships from textual context. Nevertheless, WordNet remains a valuable baseline for evaluating semantic consistency and hierarchical topic clustering within interpretable probabilistic modelling frameworks.

E. Challenges in Bridging Text Data and BHM

Integrating unstructured text data into BHM presents several critical challenges that must be addressed to enhance the effectiveness of risk analysis and decision-making processes. Understanding these challenges provides a foundation for the proposed framework outlined in the subsequent section. One of the primary challenges is data preprocessing. Unstructured text often contains significant noise and irrelevant information, which can obscure meaningful insights. Effective preprocessing is crucial for filtering out this noise and involves steps such as tokenisation, stemming, and stop-word removal. Failure to adequately preprocess text data can lead to inaccurate feature representation and difficulties in model training [23].

Another significant hurdle is feature extraction. Converting text into numerical features is a complex task that can lead to high dimensionality, which complicates model training.

Techniques such as TF-IDF and word embeddings are commonly employed. However, as the number of features increases, the model may struggle to generalise effectively, leading to overfitting and reduced performance. This phenomenon, known as the curse of dimensionality, emphasises the need for careful feature selection and dimensionality-reduction strategies to maintain a balance between feature richness and model performance [6]. Furthermore, semantic interpretation also poses difficulties. The contextual meaning of words can vary significantly, making it challenging to derive consistent interpretations. Phenomena such as polysemy (words with multiple meanings) and synonymy (different words with similar meanings) can lead to ambiguity in feature representation. This complexity necessitates advanced methods for semantic analysis to ensure the model captures the text's intended meaning [24], [25].

Establishing appropriate hierarchical structures within the BHM that accurately reflect the relationships derived from text data is another complex endeavour. The challenge lies in defining clear parameters and dependencies that are both interpretable and statistically robust[26]. This lack of clarity can hinder the model's effectiveness at capturing data traces. Moreover, the integration of topic modelling outputs with the inputs required for BHM remains largely unexplored. While prior investigations have identified effective topic models [5], linking these outputs to a BHM structure requires careful consideration of how to represent topics as factors in the model. This gap highlights the need for innovative approaches to align topic modelling outputs with BHM requirements.

Interpretability is another significant challenge. The probabilistic nature of BHMs can make it difficult to extract actionable insights from model outputs derived from unstructured text data. Ensuring that the results are interpretable and meaningful to stakeholders is essential. Thus, developing strategies to enhance interpretability is crucial for the practical use of these models [27].

Overall, existing literature demonstrates significant advances in topic modelling, semantic analysis, and Bayesian hierarchical modelling for analysing unstructured textual data. Prior studies have shown the effectiveness of topic modelling approaches in extracting latent themes from large textual corpora, while BHM provides interpretable probabilistic structures for modelling uncertainty and hierarchical relationships. However, most existing studies focus primarily on topic extraction or semantic interpretation independently, with limited attention given to systematically transforming latent textual themes into interpretable hierarchical probabilistic risk representations. Although recent advances in LLMs enable context-aware semantic interpretation, their integration within Bayesian hierarchical frameworks for business risk analysis remains underexplored.

To address this gap, this study proposes an NLP-BHM framework that integrates topic modelling, semantic labelling, and Bayesian hierarchical modelling to enable interpretable probabilistic business risk analysis from unstructured business news data.

III. Integrating Topic Modelling Outputs into Bayesian Hierarchical Models

This study presents a comprehensive framework that bridges the output from the best topic model into a Bayesian Hierarchical Model (BHM) to support probabilistic business risk modelling from unstructured business news articles. Building on prior research that identifies Non-negative Matrix Factorization (NMF) as the best-performing topic model [4], this study focuses on integrating it into the BHM structure. The framework proposes two approaches for generating interpretable topic labels, clustering related topics into higher-level factors, and modelling probabilistic relationships between these factors and associated risks. The dataset comprises 35,667 Malaysian business news articles written in English published between 2019 and 2023, collected from multiple online portals known for business-related content in Malaysia. TABLE I. provides an overview of the number of articles collected between 2019 and 2023.

TABLE I. The Number of Articles Collected from 2019-2023

Year	Number of Articles
2019	5,979
2020	6,992
2021	8,496
2022	7,824
2023	6,376

A. Framework Overview

The NLP-BHM framework operates in two phases. The first phase involves labelling topics extracted from a topic model. Each topic, represented by its top keywords, is assigned interpretable labels using two complementary strategies: LLM prompting and WordNet-based semantic analysis. The LLM approach uses OpenAI's GPT-3.5-turbo-16k to generate context-aware labels, while the WordNet approach employs semantic similarity measures and hypernym extraction to derive linguistically consistent labels. Each method is carried out independently, which allows for systematic comparison between the two approaches. The primary objective of the proposed NLP-BHM framework is to transform latent textual themes into interpretable probabilistic risk representations through hierarchical semantic structuring and Bayesian reasoning. Fig. 2 presents the proposed framework of our study.

B. Selecting Topic Model Outputs

Four topic modelling methods (LDA, NMF, Top2Vec, and BERTopic) were evaluated using coherence scores across yearly datasets from 2019 to 2023. NMF consistently achieved the highest coherence scores (0.65–0.80), indicating superior semantic consistency and interpretability for the Malaysian business news corpus. Consequently, NMF was selected as the topic-extraction component of the proposed NLP-BHM framework.

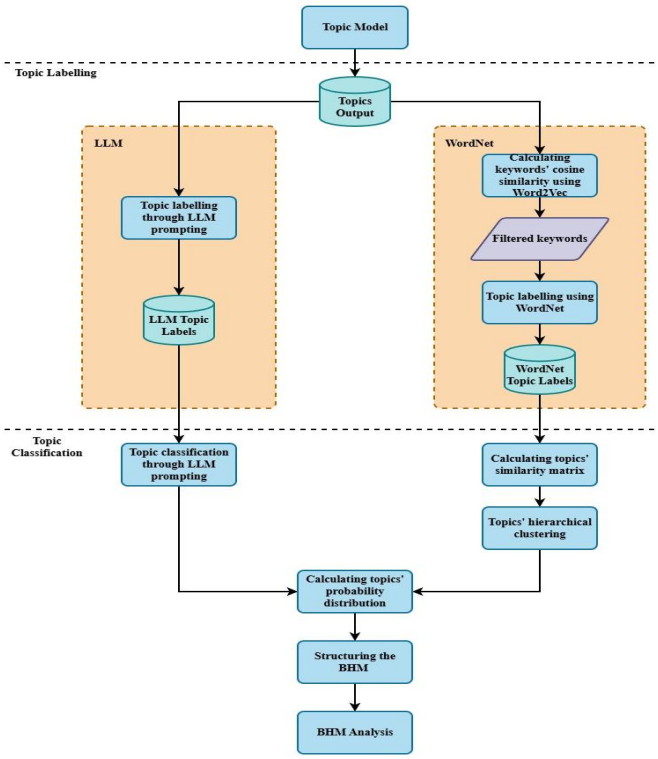


Fig. 2. An NLP-BHM framework for probabilistic business risk modelling.

C. Large Language Model Implementation

To generate interpretable topic labels, the LLM approach was employed with OpenAI's GPT-3.5-turbo-16k. For each topic extracted via NMF, the top keywords were provided to the LLM with a structured prompt instructing the model to suggest a single-word label relevant to business risk. The following prompt was used:

"I have a list of keywords as follows: {keywords}. Based on these keywords, can you suggest a single-word topic label related to risk?"

To enhance reproducibility, the API was configured with a fixed random seed, a temperature of 0.7, and a maximum token limit of 150. A second round of prompting was then applied to cluster the first-round labels into broader risk categories, using the following instruction:

"I have a list of topic labels as follows: {topic labels}. Based on these labels, can you cluster the topics together and suggest a name for each cluster?"

This two-step process helped us to transform latent topics into semantically meaningful factors and risk categories, which were then used to build the two-layer BHM structure.

D. WordNet Implementation

The WordNet-based labelling procedure involved multiple stages to ensure linguistic and semantic consistency. In the first phase, top keywords from each NMF-extracted topic were refined using cosine similarity derived from Word2Vec embeddings to eliminate semantically redundant or weakly related terms. The remaining keywords were then mapped to WordNet synsets, and their hypernyms were extracted to generate the initial topic labels. This process ensured that each label captured the core semantic meaning of its associated topic while maintaining linguistic coherence across topics. A similarity matrix was constructed from the initial topic labels using Wu-Palmer similarity, followed by hierarchical clustering with average linkage to group semantically related topics. Hypernyms were again extracted from each cluster to produce representative cluster-level labels, which were mapped as latent risks in the second layer of the BHM. This hierarchical organisation allowed the model to represent risk categories in a structured and probabilistically interpretable manner.

E. Mean Opinion Scoring Evaluation

Human evaluation remains important for assessing semantic quality and interpretability in NLP tasks, particularly where automatic evaluation metrics may not sufficiently capture contextual meaning and semantic coherence [28]. Therefore, a Mean Opinion Score (MOS) evaluation was conducted to assess the semantic clarity, contextual relevance, and interpretability of the topic labels generated by the LLM- and WordNet-based approaches. The quality of the generated labels was evaluated based on the level of understandability and interpretability. Understandability refers to how easily a label can be read and comprehended by non-specialists, while interpretability reflects how well the label captures the meaning and intent of the topic keywords in a risk-related context. Before the MOS evaluation, 20 respondents were recruited to assess the quality of the generated topic labels. The participants included a balanced representation of genders, ages, and professional statuses. This diverse composition provided a broad perspective on how different backgrounds may influence the perception of label clarity and relevance.

Each was presented with sets of topic keywords and the corresponding labels generated by the LLM- and WordNet-based approaches, while expert-assigned labels were included as a reference benchmark. Respondents were instructed to rate each label on a 5-point Likert scale, where 1 indicated a very poor fit with the topic keywords (unclear or misleading label), and 5 indicated an excellent fit (clear, relevant, and interpretable label). The evaluation was conducted via a Google Form that displayed the topic–keyword pairs alongside the candidate labels in a structured format. Respondents can enter their scores directly, ensuring consistency and ease of submission. The collected ratings were aggregated and averaged to compute the mean opinion score for each labelling method.

F. BHM Formulation

Drawing on the hierarchical architecture proposed by [4] for supply chain risk assessment, this study adopts a two-layer BHM tailored for high-dimensional textual data streams. While the framework in [4] relies on structured, predefined variables such as credit ratings and delivery metrics, our proposed model dynamically parameterises the hierarchy using latent topics (z_k) extracted from business news articles. These topics are subsequently mapped to higher-level business risk categories (r_i) through a context-aware taxonomic aggregation layer facilitated by an LLM- and WordNet-based semantic labelling. This two-layer structure allows the BHM to model uncertainty at both the thematic (topic) and categorical (risk) levels, providing the

flexibility to incorporate emerging risk factors that are typically omitted in traditional, static risk frameworks.

Let z_k denote an observed topic variable extracted from the news corpus using topic modelling, where k =1, …, K. These variables form the first layer of the hierarchy and represent the observable thematic signals derived from the textual corpus. Let r_i denote a business risk category, where i =1, …, I, represents the second layer of the hierarchy. At the topic level, the distribution of topics is represented as:

$$z_k \sim \text{Categorical}(\pi) \quad (1)$$

where, $\boldsymbol{\pi} = (\pi_1, \pi_2 \ldots, \pi_K)$ represents the topic probability distribution estimated from NMF-derived topic weights. At the risk level, each business risk category is conditionally dependent on the observed topic structure:

$$r_i \sim \text{Categorical}(\theta) \quad (2)$$

where, θ represents the probability distribution linking latent topics to higher-level business risk categories. This formulation captures how groups of latent topics contribute to higher-level business risk categories. This hierarchical design separates topic-level signals from risk-level representations, ensuring interpretability by explicitly modelling how individual topics influence higher-level business risk categories.

The probabilistic dependencies defined within the hierarchical structure are parameterised using topic-weight distributions and semantic associations derived from the topic labelling process. Following the Bayesian aggregation strategy proposed in [4], probability estimates are propagated through the hierarchy to quantify the relative influence of latent topics on higher-level business risk categories.

G. Obtaining Prior Probabilities

Initially, the NMF model produced weights for each keyword within a topic, representing the strength of the keyword's contribution to that topic. For each topic, the mean keyword weight was computed and subsequently used as the first-level probability estimate within the Bayesian structure. Following the strategy adopted in [4], these mean topic weights serve as the prior probability estimates for the corresponding latent topics. The prior probability for topic z_k is calculated as:

$$\pi_k = \frac{1}{n_k} \sum_{j=1}^{n_k} W_{kj} \quad (3)$$

where, W_{kj} represents the NMF weight of each keyword j within topic k,

n_k is the total number of keywords within topic k,

and π_k represents the prior probability estimate of topic k.

H. Parameter Estimation and Bayesian Risk Aggregation

The topic probability distribution π was estimated using the mean NMF keyword weights obtained from Eq. (3). Each topic probability reflects the relative prevalence of a latent topic within the business news corpus and serves as the initial probabilistic representation at the factor level of the hierarchy. The relationship between latent topics and business risk categories was established through semantic clustering generated using the LLM- and WordNet-based labelling approaches. Topics assigned to the same semantic cluster were grouped to form higher-level business risk categories.

Following the hierarchical aggregation approach proposed in [4], the probability of a business risk category, r_i was estimated by aggregating the probabilities of its associated topics:

$$P(r_i) = \frac{\sum_{k=1}^{m} \pi_k}{m} \quad (4)$$

where:

- π_k denotes the probability of topic k,
- m denotes the number of topics associated with risk category r_i.

Consistent with [4], the contribution weight associated with topic k was assumed to be equal to 1. This aggregation produces a probability estimate for each business risk category by combining the contributions of semantically related latent topics. Subsequently, the overall Business Risk Impact (BRI) score was computed by aggregating the estimated probabilities of all business risk categories, again assuming the contribution weight assigned to the risk category is equal to 1:

$$BRI = \frac{\sum_{i=1}^{I} P(r_i)}{I} \quad (5)$$

where:

- $P(r_i)$ denotes the probability of business risk categories i,
- I denote the total number of business risk categories.

The resulting BRI score provides a comparative measure of the overall business risk represented within each annual news corpus. To maintain computational efficiency and interpretability, probability estimates were obtained through hierarchical aggregation of topic probabilities and semantic cluster associations, following the Bayesian hierarchical framework in [4]. The resulting probability estimates form the basis of the comparative business risk analysis presented in the Results section.

IV. Results and Discussion

This section presents a comparison of LLM- and WordNet-based labelling integrated into a BHM. The results are organised to demonstrate each method's contribution to the NLP-BHM framework. Due to space constraints, the final BHM structures for the 2020 and 2021 data are illustrated in the figures.

A. Comparison of LLM and WordNet Labelling Effectiveness Through Mean Opinion Scoring

Fig. 3 shows the MOS results for 2019 topics generated by LLM and identified by WordNet.

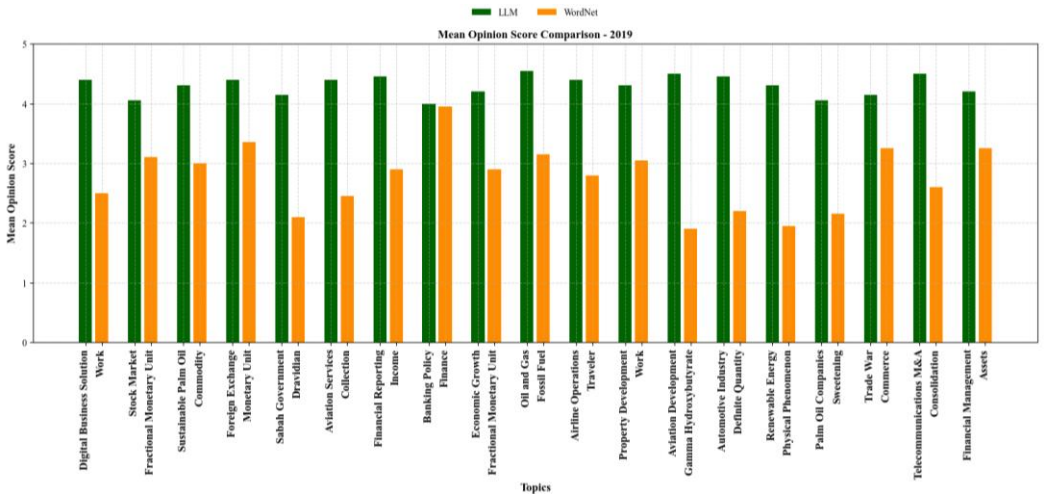


Fig. 3. Mean Opinion Scores from respondents to vote on LLM and WordNet topics from 2019 data.

The results indicated that respondents consistently favoured LLM labels (dark green), which outperformed the WordNet-based labels (orange). The LLM labels were rated as more understandable and better aligned with the business risk context. In contrast, WordNet labels were often perceived as too broad or disconnected from the underlying risk themes. Similar MOS results are observed in the 2020 to 2023 data, consistent with the 2019 data and highlighting patterns favouring LLM labels over WordNet labels. TABLE TABLE II. shows the average MOS results for 2019-2023 topics.

TABLE II. AVERAGE MOS SCORES FOR THE TOPICS GENERATED BY LLM AND IDENTIFIED BY WORDNET

Year	Average MOS (LLM)	Average MOS (WordNet)
2019	4.303	2.766
2020	4.288	2.683
2021	4.358	2.721
2022	4.369	2.634
2023	4.316	2.729

Over the five years, the average MOS for LLM topics increased from 4.2 to 4.4, reflecting a growing appreciation for their clarity and relevance. Conversely, WordNet topics maintained lower scores, ranging from 2.6 to 2.8, as they were often seen as too broad and less connected to specific risks. This trend emphasises the effectiveness of LLMs in enhancing topic interpretation and relevance in business contexts.

B. LLM-Based Labelling and Clustering Results

The factors and risks derived from the LLM topics were initially assigned prior probabilities based on the topics' weights from the NMF topic model. These priors represent the baseline importance of each factor before considering any risk evidence. Using the BHM, these priors were updated into posterior probabilities, reflecting the likelihood of specific risks once related factors are considered. The BHM structures for 2020 and 2021, as depicted in Fig. 4 and Fig. 5, reveal evolving business risk landscapes. In 2020, the focus was on "Transportation and Travel" and "Finance and Economy," with finance-related topics heavily influencing risk perception. By 2021, attention shifted to "Industry Analysis" and "Business and Finance," indicating a heightened emphasis on sectors like aviation and automotive. The business risk impact score increased from 0.408 to 0.444, reflecting heightened concerns, particularly in finance and governance. The reported risk impact scores represent relative probabilistic risk representations derived from posterior topic–risk associations within the BHM framework, rather than direct forecasts of future business outcomes. This evolution highlights shifting priorities and perceived risks in the business environment over the years.

Similar thematic transitions were observed across the 2019, 2022, and 2023 datasets, reflecting evolving business priorities and emerging economic concerns over time. TABLE III. shows the overall business risk impact score obtained from the LLM-based BHM structuring.

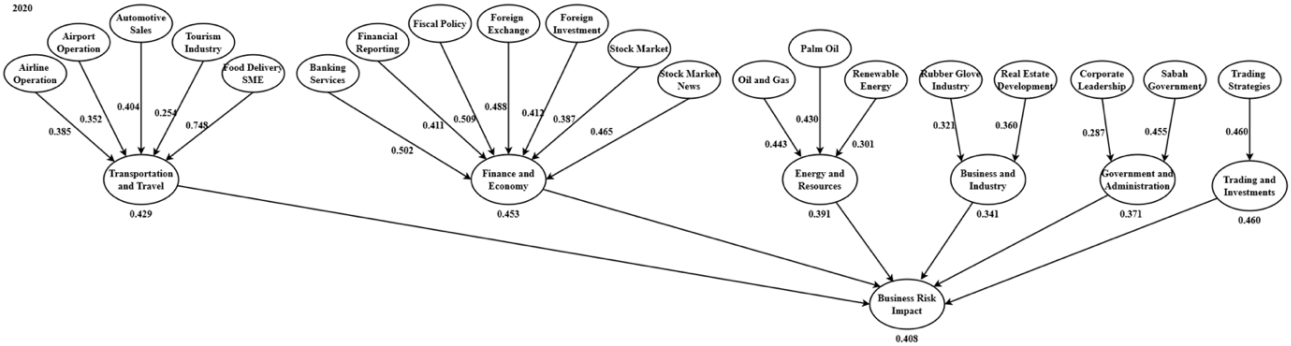


Fig. 4. LLM-based BHM structure for 2020 business news data.

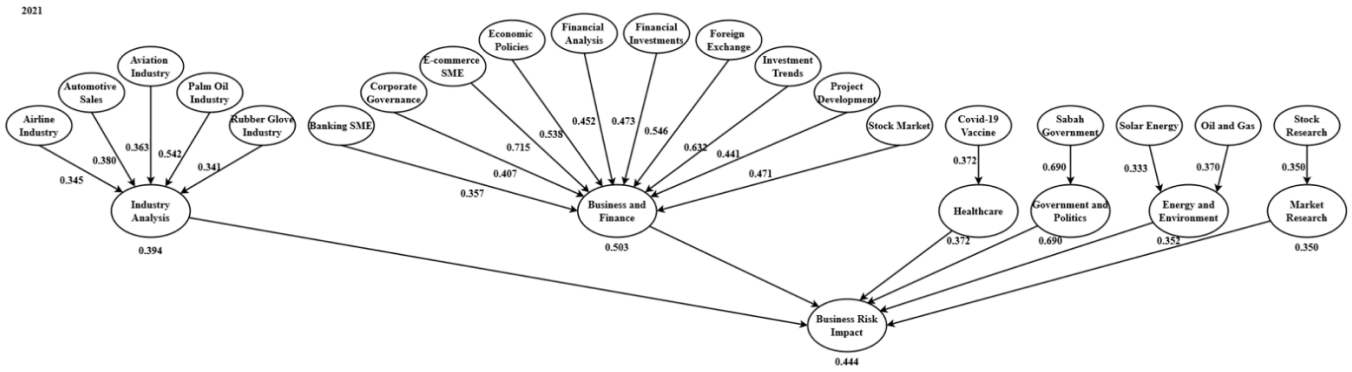


Fig. 5. LLM-based BHM structure for 2021 business news data.

TABLE III. OVERALL BUSINESS RISK IMPACT SCORE FOR LLM-BASED BHM STRUCTURE

Year	Business Risk Impact
2019	0.492
2020	0.408
2021	0.444
2022	0.430
2023	0.458

As shown in Table III, the LLM-based BHM structure produced business risk impact scores ranging from 0.408 to 0.492 across the study period. The highest risk impact score was observed in 2019 (0.492), followed by a decline in 2020 (0.408), before gradually increasing to 0.444 in 2021, 0.430 in 2022, and 0.458 in 2023. These comparisons illustrate evolving business priorities, reflecting changes in the economic landscape and policy environment over the years. These results highlight the NLP-BHM's ability to capture both persistent and emerging risks in Malaysia's business landscape.

C. WordNet-Based Labelling and Clustering Results

The WordNet-based Wu–Palmer similarity scores were used to construct a semantic similarity matrix representing the conceptual relatedness among topic labels. Higher similarity values indicated stronger semantic associations between topics, while lower values reflected weaker conceptual relationships. The similarity matrix was subsequently transformed into a distance matrix to support hierarchical clustering using the average-linkage method.

Fig. 6 depicts the resulting dendrogram for the sample, which highlights the semantic organisation of the topics in 2020. The height of each merge reflects the degree of dissimilarity between clusters, allowing for interpretation of how closely or distantly related the topics are. By applying a predefined threshold of 0.7, distinct semantic groups are identified, enabling clearer interpretation of the underlying conceptual structure derived from WordNet. It is observed that in that particular year, the topics were divided into four groups. After WordNet-labelled labels were processed through a similarity matrix and hierarchical clustering to group the factors into appropriate risk labels, the BHM's priors were estimated to reflect how common each factor was in the data, and then posteriors were calculated for the risks associated with each factor.

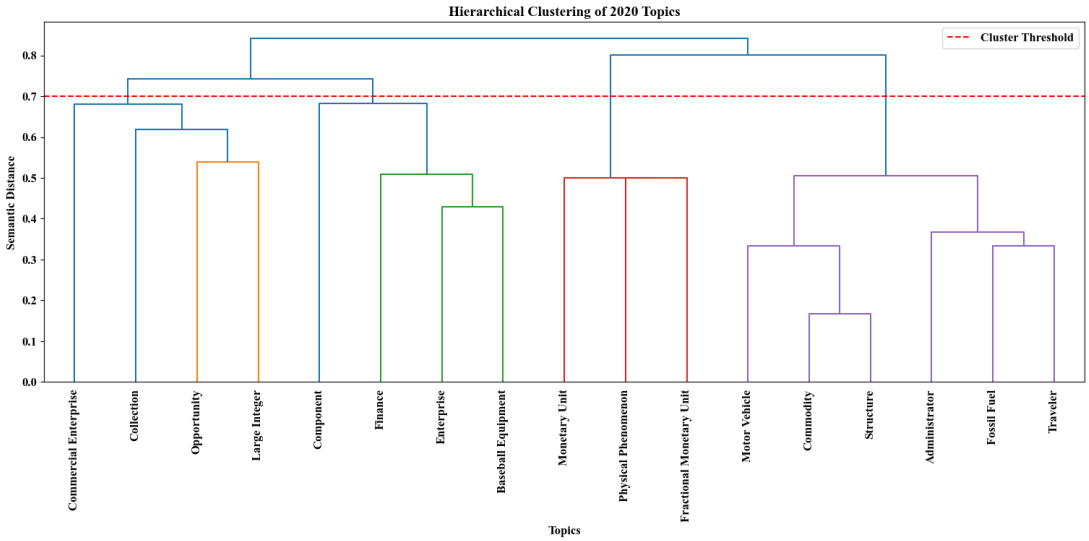


Fig. 6. Result of hierarchical clustering for topics in the 2020 data.

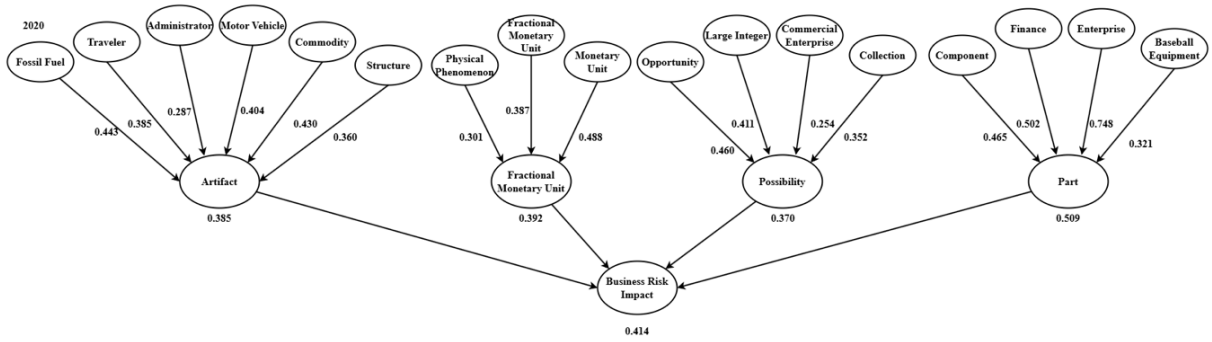


Fig. 7. WordNet-based BHM structure for 2020 business news data.

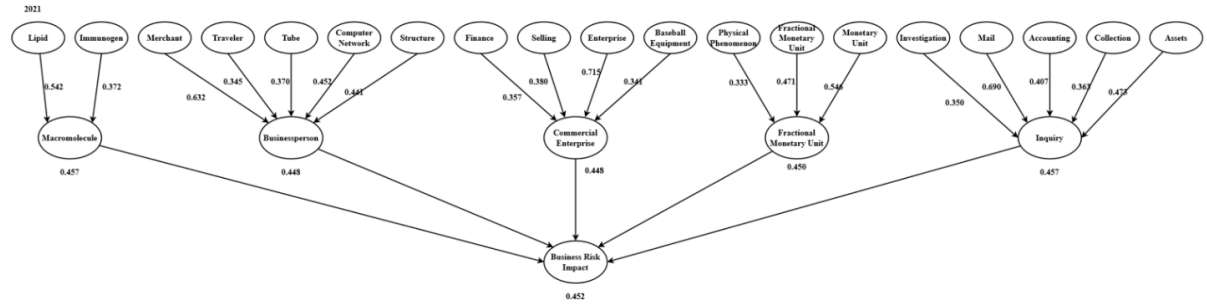


Fig. 8. WordNet-based BHM structure for 2021 business news data.

Fig. 7 and Fig. 8 show the resulting 2020 and 2021 BHM structures, respectively, obtained by leveraging WordNet-based labelling. The WordNet-based structures for 2020 and 2021 reveal distinct thematic focuses on business risk analysis. In 2020, the structure highlights categories such as "Artifact" and "Part," emphasising elements like fossil fuels, commodities, and financial components, leading to a business risk impact of 0.414. By 2021, the focus shifts to "Businessperson" and "Inquiry," with themes involving finance, enterprise, and accounting, reflecting a comprehensive approach to commercial enterprise risks with an impact score of 0.452. These structures illustrate evolving priorities in risk management, with a transition from tangible assets to more abstract business and financial concerns. Similar semantic transitions were observed across the remaining yearly datasets, where the WordNet-based structures progressively shifted from tangible asset-oriented concepts toward broader business and financial categories. TABLE IV. shows the overall business risk impact score obtained from the WordNet-based BHM structuring.

TABLE IV. OVERALL BUSINESS RISK IMPACT SCORE FOR WORDNET-BASED BHM STRUCTURE

Year	**Business Risk Impact**
2019	0.423
2020	0.414
2021	0.452
2022	0.416
2023	0.451

As shown in Table IV, the WordNet-based BHM structure produced business risk impact scores ranging from 0.414 to 0.452 over the five-year study period. The highest risk impact score was recorded in 2021 (0.452), followed closely by 2023 (0.451), while the lowest score was observed in 2020 (0.414). The remaining years, 2019 and 2022, yielded scores of 0.423 and 0.416, respectively. The variation in risk impact scores across the five years is relatively small, with a difference of only 0.038 between the highest and lowest values.

D. Discussion

The proposed NLP-BHM framework represents an integrated approach that combines topic modelling, BHM, and language model techniques to transform unstructured business news into structured, actionable risk insights. By integrating probabilistic reasoning with semantic enrichment, the framework addresses the complexity and uncertainty inherent in business risk analysis, enabling a more structured interpretation of latent risk factors. A key strength of the framework is the dynamic nature of the BHM, which allows risk probabilities to be continuously updated as new information becomes available. This makes the framework suitable for fast-changing business environments and supports near real-time risk assessment and decision-making.

The hierarchical structure of the model further improves interpretability by explicitly modelling relationships between high-level risks and underlying sub-factors, allowing clearer tracing of dependencies within complex risk systems. In addition, the use of LLM-based topic labelling enhances semantic representation by producing contextually meaningful, human-interpretable labels, thereby improving the usability of extracted risk topics. Based on the MOS results, LLM-based labelling is perceived as more interpretable than WordNet-based labelling, indicating its strength in capturing richer semantic context in business news. Importantly, in this study, LLM

outputs are made consistent through a fixed random seed, ensuring reproducibility and addressing concerns regarding variability in language model generation. In contrast, WordNet-based labelling provides deterministic, consistent labels that support reproducibility and standardisation. However, its limited coverage of domain-specific vocabulary, particularly in business and financial contexts, often leads to overly general topic labels that reduce interpretability and granularity. As a result, WordNet is less effective in distinguishing fine-grained risk categories, although it remains useful as a stable baseline for structured comparison.

It should be noted that the resulting business risk impact scores represent relative risk indicators derived from textual topic distributions rather than direct measures of real-world business risk severity. The framework quantifies the prominence of risk-related themes within the news corpus and enables comparative analysis across different periods. As such, higher scores indicate a greater concentration of risk-related topics within the analysed news data, rather than a verified increase in actual business risk outcomes. The current framework terminates at the estimation of business risk impact. However, these probabilistic indicators could be further served as inputs to more advanced risk quantification and decision-support models. Similar to the approach adopted in [4], where disruption probabilities were further utilised to estimate supplier revenue impact and Value at Risk (VaR), the proposed business risk impact scores could potentially be integrated with financial, operational, or market-based indicators to quantify the potential consequences of emerging business risks. Consequently, the business risk impact score should be viewed not only as a comparative textual risk indicator but also as a foundation for future quantitative risk assessment frameworks that link news-derived risk signals with real-world business outcomes.

V. Conclusion and Future Work

This study proposed an NLP-BHM framework for transforming topic modelling outputs into interpretable hierarchical probabilistic representations for business risk analysis. Using a corpus of Malaysian business news articles published between 2019 and 2023, four topic modelling approaches were evaluated, where NMF consistently produced the most coherent and semantically interpretable topic structures. The framework subsequently integrated semantic topic labelling and clustering using both LLM-based prompting and WordNet-based semantic analysis within a Bayesian hierarchical modelling structure.

The study demonstrates that topic modelling outputs can be effectively transformed into interpretable hierarchical probabilistic risk representations through semantic labelling and Bayesian hierarchical structuring, thereby addressing the challenge of integrating unstructured textual themes into uncertainty-aware business risk analysis. Comparative findings further show that LLM-based semantic labelling generated more contextually relevant and interpretable topic representations than the WordNet-based approach, as reflected by consistently higher MOS scores. While WordNet-based clustering provided semantically structured and reproducible groupings, its limited contextual flexibility reduced its effectiveness in capturing domain-specific business terminology and evolving semantic relationships.

Overall, the proposed NLP-BHM framework provides a structured approach to bridging latent textual themes and higher-level probabilistic risk representations through hierarchical semantic structuring and Bayesian reasoning. By integrating topic modelling, semantic interpretation, and posterior risk representation within a unified framework, the study demonstrates the potential of NLP-driven probabilistic analysis for interpretable business intelligence and decision support.

Future work may explore integrating dynamic temporal Bayesian structures, automated probabilistic parameter learning, and multiple LLMs to evaluate the generalisation ability, robustness, and consistency of semantic topic labelling across different model architectures. In addition, further research should investigate the validation of the generated business risk indicators against external economic and financial benchmarks, such as market volatility indices, business confidence indicators, or observed business risk events, to provide stronger empirical evidence for interpreting the resulting risk impact scores as meaningful measures of real-world business risk severity and to enhance the practical applicability of the proposed NLP-BHM framework.

Acknowledgment

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT06/UNIMAS/02/2). The authors thank the MOS participants. The survey was approved by the UNIMAS Human Research Ethics Committee (HREC(NM)/2023(2)/57).

Declaration on Generative AI

This work utilises OpenAI's ChatGPT to enhance data analysis and interpretation. All outputs generated through the technology have been critically reviewed and validated by the authors to ensure accuracy and integrity.

References

[1] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Frontiers in Sociology, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

[2] G. M. Allenby, P. E. Rossi, and R. E. Mcculloch, "Hierarchical Bayes Models: A Practitioners Guide Hierarchical Bayes Models," 2005.

[3] M. Veenman, A. M. Stefan, and J. M. Haaf, "Bayesian hierarchical modeling: an introduction and reassessment," Behav. Res. Methods, May 2023, doi: 10.3758/s13428-023-02204-3.

[4] S. M. Ali, A. B. M. M. Bari, A. A. M. Rifat, M. Alharbi, S. Choudhary, and S. Luthra, "Modelling supply chain disruption analytics under insufficient data: A decision support system based on Bayesian hierarchical approach," International Journal of Information Management Data Insights, vol. 2, no. 2, Nov. 2022, doi: 10.1016/j.jjimei.2022.100121.

[5] V. Asas, S. S. Juan, V. C. Kerbun, S. Chua, J. Labadin, and E. Lau, "A Comparative Analysis of Topic Modelling Techniques for Malaysian Business News Data: LDA, NMF, Top2Vec, and BERTopic," in 2025 14th International Conference on Information Technology in Asia (CITA), IEEE, Aug. 2025, pp. 42–47. doi: 10.1109/CITA66455.2025.11198734.

[6] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.

[7] L. Kong and A. Brintrup, "A hierarchical Bayesian model for payment delay prediction in supply chain financing," Int. J. Prod. Res., pp. 1–24, Oct. 2025, doi: 10.1080/00207543.2025.2546029.

[8] M. Coll, M. Grazia Pennino, J. Steenbeek, J. Sole, and J. M. Bellido, "Predicting marine species distributions: Complementarity of food-web and Bayesian hierarchical modelling approaches," Ecol. Modell., vol. 405, pp. 86–101, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.05.005.

[9] A. Couture et al., "Estimating COVID-19 Hospitalizations in the United States with Surveillance Data Using a Bayesian Hierarchical Model: Modeling Study," JMIR Public Health Surveill., vol. 8, no. 6, 2022, doi: 10.2196/34296.

[10] H. Zeng et al., "Climate-informed clustering based nonstationary regional extreme flood events spatio-temporal evolution using hierarchical Bayesian modeling," J. Hydrol. Reg. Stud., vol. 56, Dec. 2024, doi: 10.1016/j.ejrh.2024.102066.

[11] A. Talman, H. Celikkanat, S. Virpioja, M. Heinonen, and J. Tiedemann, "Uncertainty-Aware Natural Language Inference with Stochastic Weight Averaging," in Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), T. Alumäe and M. Fishel, Eds., Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 358–365. [Online]. Available: https://aclanthology.org/2023.nodalida-1.37/

[12] D. Yu and B. Xiang, "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling," Expert Syst. Appl., vol. 225, Sep. 2023, doi: 10.1016/j.eswa.2023.120114.

[13] R. Vangara et al., "Finding the Number of Latent Topics With Semantic Non-Negative Matrix Factorization," IEEE Access, vol. 9, pp. 117217–117231, 2021, doi: 10.1109/ACCESS.2021.3106879.

[14] D. Angelov, "Top2Vec: Distributed Representations of Topics," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.09470

[15] S. S. Pangastuti, E. N. Rohmatullayaly, and N. Najmi, "TOPIC MODELING FOR USER FEEDBACK DATASET," Communications in Mathematical Biology and Neuroscience, vol. 2025, 2025, doi: 10.28919/cmbn/8932.

[16] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[17] O. Lezhnina, "Depression, anxiety, and burnout in academia: topic modeling of PubMed abstracts," Front. Res. Metr. Anal., vol. 8, 2023, doi: 10.3389/frma.2023.1271385.

[18] I. Widiastuti and H. S. Yong, "TR-GPT-CF: A Topic Refinement Method Using GPT and Coherence Filtering," Applied Sciences (Switzerland), vol. 15, no. 4, Feb. 2025, doi: 10.3390/app15041962.

[19] S. Minaee et al., "Large Language Models: A Survey," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2402.06196

[20] Y. Li, S. Wang, H. Ding, and H. Chen, "Large Language Models in Finance: A Survey," in ICAIF 2023 - 4th ACM International Conference on AI in Finance, Association for Computing Machinery, Inc, Nov. 2023, pp. 374–382. doi: 10.1145/3604237.3626869.

[21] G. A. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[22] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," in 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA: Association for Computational Linguistics, Jun. 1994, pp. 133–138. doi: 10.3115/981732.981751.

[23] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[24] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, 3rd ed. 2026. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[25] D. Jurafsky and J. H. Martin, "Speech and Language Processing (2008)," vol. 1, 2008, doi: 10.1162/089120100750105975.

[26] R. Navigli, "Is Word Sense Disambiguation Dead in the LLM Era?," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 46, pp. 39753–39762, Mar. 2026, doi: 10.1609/aaai.v40i46.41331.

[27] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv: Machine Learning, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:11319376

[28] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "BLEU is Not Suitable for the Evaluation of Text Simplification," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 936–941

Artificial Intelligence and the Transformation of Academic Integrity in Higher Education: A Systematic Review

Wannakorn Phornprasert[1], Wongpanya S. Nuankaew[2], Pratya Nuankaew[3*]
Department of Educational Measurement and Evaluation-School of Education,
University of Phayao, Phayao, 56000 Thailand[1]
Department of Computer Science-School of Information and Communication Technology,
University of Phayao, Phayao, 56000 Thailand[2]
Department of Digital Business-School of Information and Communication Technology,
University of Phayao, Phayao, 56000 Thailand[3]

***Abstract*—This study examines how Artificial Intelligence (AI) is transforming academic integrity in higher education, altering both learning opportunities and the risks associated with misconduct. As creative AI tools become embedded in everyday academic work, they provide valuable support for writing, research assistance, and skills development. Still, they also challenge long-held assumptions about authorship, originality, and assessment. Emerging evidence suggests that students are using AI in a variety of ways, from supporting legitimate learning to producing fully automated assignments. However, AI-driven integrity technologies, such as plagiarism detectors and authorship checking models, are becoming more effective but continue to face issues of bias, false positives, and limited transparency. This rapid shift has created a gap between technological change and academic readiness, highlighting the need for institutions to rethink assessment design, improve integrity frameworks, and foster a culture of responsible AI use, rather than relying solely on surveillance and sanctions. This review compiles the latest studies published between 2020 and 2025 to map current practices, risks, and policy responses. The findings suggest that academic integrity in the age of artificial intelligence (AI) cannot be focused solely on preventing fraud. But this needs to expand to support ethical digital literacy, redesign learning tasks that require human reasoning, and ensure fairness in automated decision-making systems. The study concludes with recommendations for educators, researchers, and policymakers to balance innovation with responsibility to ensure that AI becomes a tool for transforming learning, rather than a threat to academic values.**

***Keywords*—AI in education; academic integrity; Generative AI; ethical digital literacy; assessment transformation**

I. Introduction

The usage of Artificial Intelligence (AI) in education has changed the education landscape by improving personalized learning, such as customized learning pathways by adapting to the different needs and learning styles of individual students [1]. This makes administrative tasks more efficient through methods such as automated grading, which allow teachers to focus more on teaching and improving student engagement [2]. However, the usage of AI in education also posts ethical concerns and challenges, such as data privacy [3], the quality of the data used, and the need for ongoing training for teachers and students [1].

A. Background

AI has become more recognized as a personalized learning tool in education. AI-driven platforms allow the analyzation of students' performance data, creating a customized learning path and addressing individual strengths and weaknesses [4]. This feature allows students to self-regulate their learning and take control of their own education journey [5]. Additionally, AI has also become more recognized for its potential in increasing efficiency by automating assessment and feedback in education. Tools such as Automated Essay Scoring (AES) reduce administrative burden on teachers by enabling fast and consistent grading [6]. Moreover, AI has also been used as a tool to support students' success. AI tools such as chatbots offer a personalized self-service platform, where students can address any question and get help with course selection in education [7].

The integration of Generative AI (GAI) gives rise to an opportunity for enhancing learning by offering a personalized learning journey to students, this foster engagement and critical thinking in students [8]. These features help in students' comprehension of the course materials [9]. However, this also gives rise to a challenge in academic integrity. Due to the ease of generating content, there are risks such as dishonesty, plagiarism [10], and even a decline of cognitive skills and critical thinking from overreliance on GAI [11].

In higher education, AI assists students through the inclusion and enhancement of learning by personalization, which makes education more accessible to students with diverse needs [12]. Those with disabilities or those facing language barriers can benefit from these AI features, as AI allows them to learn from anywhere and generate content in their preferred language [13]. In teaching assistant and virtual classroom, AI tools provide personalized study plans and support that help in explaining complex concepts [14]. Additionally, AI creates an immersive learning environment with the uses of things such as virtual reality, which provide a safe space for students to practice their skills [15].

*Corresponding author

B. Research Objectives

The primary objective of this research is to conduct a systematic review of AI applications in higher education across seven key topics by searching for recent literature to identify both opportunities for enhancing learning experiences and associated challenges. This study aims to evaluate how AI tools contribute to personalized learning, automated assessment, academic advising, academic integrity, inclusion, virtual teaching support, and institutional readiness while maintaining unbiased, keeping ethical considerations in mind, and considering practical implications for teachers and students.

C. Scope and Significance

This review focuses on peer-reviewed studies published between 2020 and 2025, emphasizing AI's role in higher education within the seven specified topics. It excludes non-academic sources and broader K-12 applications to maintain a targeted analysis. The significance lies in providing teachers, administrators, and policymakers with a concise framework for understanding AI's transformative potential, addressing gaps in ethical integration, and fostering informed strategies to improve student outcomes and institutional efficiency in an evolving educational landscape.

D. Structure Overview

This study is structured as follows: Section I introduces the topic, background, objectives, scope, and significance. Section II outlines the proposed systematic review method and framework. Section III details the methodology of implementation, search processes, quality assessment and limitations. Section IV presents results and discussions for each of the seven topics in Section V. Section VI concludes with key findings and recommendations for future research.

II. Proposed Methodology

A. Systematic Review Approach

The systematic review method used in this study was designed to provide a comprehensive, transparent, and replicable way to synthesize existing research on the use of Artificial Intelligence (AI) in higher education. This approach ensures that the review maintains strong methodological standards while reducing bias in how studies are selected, evaluated, and interpreted. It also helps identify trends, challenges, and research gaps across different academic settings related to AI's integration in higher education.

The review follows the principles of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, which offers structured guidance for finding, screening, and reporting research evidence. By following this framework, the study guarantees that each step—from searching databases to final inclusion—is transparent and well-documented. The review focuses on seven key themes that represent critical areas of AI use in higher education: personalized learning, automated assessment, academic advising, academic integrity, inclusion, virtual teaching support, and institutional readiness.

The process started with a systematic search across selected academic databases, including IEEE Xplore, Google Scholar, and Scopus, chosen for their wide coverage of peer-reviewed research in educational technology and AI applications. The search combined relevant keywords and Boolean operators to find a broad range of studies published from 2020 to 2025. This timeframe was selected to reflect recent advances in AI and to match the current era of technological change shaping higher education.

Each study found was reviewed through a structured screening process in two stages: an initial review of titles and abstracts, then a detailed examination of the full texts to determine eligibility. Inclusion criteria specified that studies focus on higher education, present empirical or conceptual analyses of AI tools, and be published in peer-reviewed journals. Materials that were not academic, such as editorials or studies unrelated to higher education, were excluded. After selecting eligible studies, all were assessed for quality using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) framework. This ensured that the evidence included in the review was credible, relevant, and methodologically strong. The final dataset supported an analysis of how AI is transforming teaching, learning, and institutional operations within higher education.

Overall, this systematic review method provided a structured, evidence-based way to explore AI's transformative effects. It highlights both potential benefits and the ethical, pedagogical, and institutional challenges that come with AI integration. This approach ensures that the conclusions are based on reliable, diverse academic evidence.

B. Proposed Framework

The proposed framework for this systematic review was developed to ensure a transparent, reproducible, and unbiased process when analyzing existing literature on artificial intelligence (AI) applications in higher education. It provides a clear and structured pathway for identifying, selecting, and synthesizing relevant studies, helping readers understand the methodology and reasoning behind each step of the review. Essentially, the framework follows a predefined, step-by-step protocol inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. This protocol emphasizes consistency, accuracy, and traceability throughout the review process. It consists of three main stages: searching, selecting, and synthesizing.

In the searching stage, relevant academic databases — IEEE Xplore, Google Scholar, and Scopus — were systematically explored to find peer-reviewed publications from 2020 to 2025. This time frame was chosen to focus on recent developments and emerging practices in AI, reflecting the fast-changing nature of technology and education. The search strategy used a combination of specific keywords related to the seven themes: personalized learning, automated assessment, academic advising, academic integrity, inclusion, virtual teaching support, and institutional readiness.

The selection stage involved a multi-step screening process to reduce subjectivity and bias. First, duplicates and non-academic sources were removed. Next, studies were filtered based on relevance of title and abstract, followed by a full-text review to confirm they met the inclusion criteria—specifically,

that each study involved AI applications in higher education and was published in a peer-reviewed outlet.

Finally, in the synthesizing stage, all eligible studies were analyzed and grouped according to the seven key areas. Data was extracted using a structured template to record essential details such as research goals, methodologies, AI techniques employed, main findings, and limitations. The synthesis aimed to combine findings across studies to identify common patterns, emerging themes, and significant challenges in the field.

This systematic framework offers a balanced and comprehensive approach for understanding the opportunities and challenges of AI in higher education. By applying consistent criteria, selecting high-quality sources, and maintaining transparency, the review minimizes bias and ensures that the analysis provides both clarity and credibility in illustrating the evolving relationship between AI and higher education practices.

C. Visualization

This systematic review process is for identifying and selecting studies for a quality system. This review highlights PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) benefits, such as providing an intuitive abstraction layer that simplifies database operations and improves code quality, security, and scalability [51, 52].

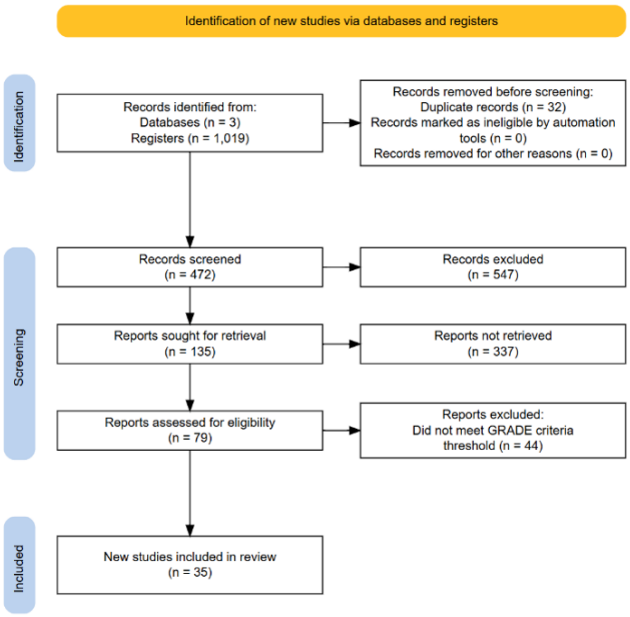


Fig. 1. PRISMA flowchart for inclusion and exclusion of works.

The PRISMA flow diagram in Fig. 1 shows the review process in the beginning, with 1,019 studies identified from 3 databases. Prior to screening, 32 duplicate records and ineligible records marked by automation tools were removed, reducing the total to 472 records screened. From these, 547 records were excluded, and then 135 reports were sought for retrieval. Of the retrieved reports, 79 were assessed for eligibility, excluding 44 that did not meet GRADE criteria, resulting in 35 new studies being included in the review.

D. Expectations

The review expects to produce clear summaries for each topic and to highlight both benefits and challenges. The process and result of these summaries will determine the methodology and results, which will provide insights for teachers as a clear, concise, systematic review structure.

III. Methodology

A. Implementation

The implementation phase of this systematic review was conducted by strictly following the proposed framework to ensure methodological rigor, transparency, and reproducibility. The process began with an extensive search across major academic databases, namely IEEE Xplore, Google Scholar, and Scopus, which were chosen for their vast collections of peer-reviewed research in education, technology, and artificial intelligence. The search strategy used a combination of predefined keywords and Boolean operators designed to find relevant literature on seven key themes: personalized learning, automated assessment, academic advising, academic integrity, inclusion, virtual teaching support, and institutional readiness.

The data collection was limited to the period between 2020 and 2025 to reflect the latest developments in AI-driven educational tools and to ensure that the review addressed the current technological and pedagogical landscape. Search results were first screened by titles and abstracts to remove unrelated or duplicate studies. Then, full-text articles were reviewed to verify eligibility based on criteria such as publication in peer-reviewed journals, relevance to higher education, and direct discussion of AI applications within one or more of the seven focus areas.

Once selected, data from each eligible study were systematically extracted using a standardized data extraction form. This form recorded key details, including study goals, research methods, AI technologies used, outcomes, limitations, and implications for higher education. The extraction aimed to ensure consistency across studies while keeping contextual details for accurate synthesis.

To enhance the reliability and validity of the combined findings, each study underwent a structured quality assessment using the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) framework. This assessment focused on criteria such as methodological quality, risk of bias, relevance to the research questions, and clarity of the presented evidence. Studies rated as “high” or “moderate” in quality were prioritized in the synthesis to increase the trustworthiness of the conclusions.

Finally, the collected and evaluated data were organized into clear thematic summaries aligned with each of the seven topics. The synthesis highlighted patterns and interpreted differences across studies. The findings were presented in clear, accessible academic language to provide educators, researchers, and policymakers with insights into the roles and challenges of AI in higher education. This systematic approach not only helped organize and validate diverse research findings but also built a transparent basis for further analysis and discussion within the review.

B. Search and Selection Process

The search uses databases IEEE, Google Scholar, and Scopus, which contain numerous peer-reviewed records that are regularly updated. Studies criteria, such as peer-reviewed and dated between 2020 and 2025, were included while also following the PRISMA process to ensure transparency and to manage the scope within the study's limits.

C. Quality Assessment

The quality of selected studies was evaluated using the GRADE (Grading of Recommendations, Assessment, Development and Evaluations) framework to ensure reliable evidence for the systematic review of AI in higher education. GRADE assesses studies based on criteria like risk of bias, inconsistency, imprecision, and indirectness by assigning each criterion a quality rating (high, moderate, or low). For this research, studies were examined for relevance and methodological clarity to the seven topics. This streamlined process ensured only credible sources shaped the synthesis to maintain the focus and transparency of this study [53, 54].

D. Limitations

This systematic review is constrained by its reliance on selected databases (IEEE, Google Scholar, and Scopus), which may overlook relevant studies in other repositories, like non-English publications. The time frame of 2020-2025 limits the historical context, which may also potentially miss foundational developments in AI applications. The streamlined quality assessment using GRADE may introduce subjective judgments, and the focus on peer-reviewed sources may exclude emerging gray literature, which could affect the overall insight into the rapidly evolving AI technologies in higher education.

IV. Results

A. AI as a Pedagogical Tool: Personalized and Adaptive Learning

Artificial Intelligence (AI) transforms education by enabling personalized adaptive learning experiences. Traditional teaching methods come with fixed pacing and normal content, while AI systems customize instruction to each student's progress preferences and performance data. AI tools can offer immediate feedback and targeted remediation, which allows students to learn at their own speed [16]. These tools excel in subjects like math and science, where student abilities often differ greatly. By analyzing learner behavior in real-time, AI helps teachers meet diverse needs in large classrooms and improve learning outcomes [17]. AI tutoring systems and virtual assistants enhance personalization by providing real-time support and recommending resources tailored to students' actions. These systems help with customized content, guiding students through problem-solving and constant feedback [17]. Using these systems with personalized feedback boosts comprehension and increases student motivation and emotional engagement [18]. These systems can also be paired with certain environments, like AI-extended reality, by adapting tasks based on student performance and supporting the students using experiential learning to further help with student and system development [19].

However, the use of AI in teaching can cause some ethical and professional concerns. Automation can allow for streamlining of tasks like grading work and feedback, but risks alienation between students and teachers, especially if the work is streamlined, the grading or feedback may lack the correct context [20]. Decisions like this may limit the teachers' ability and the students' experiences. The alternative is to use models that require human management or input during critical decision-making processes, which will ensure that teachers use technology as a supportive tool rather than a replacement, allowing for the adjustment of features of the AI and giving students an overall good experience [20].

TABLE I. AI IN PERSONALIZED AND ADAPTIVE LEARNING: SUMMARY AND GRADE ASSESSMENT

No.	Key Points	Challenges	Effect on Teachers and Students	GRADE
[16]	Tools like Knewton and ALEKS can provide immediate and targeted feedback to improve results.	Automation risks alienating students from teachers due to a lack of context in feedback.	Teachers use AI as a supportive tool while students gain motivation through personalized feedback.	Moderate
[17]	AI analyzes learner behavior in real time for customized content and problem-solving guidance.	Decisions may limit teacher autonomy if not human-managed.	Boosts comprehension and emotional engagement to help close achievement gaps in diverse classrooms.	Moderate
[18]	System tracks emotional and cognitive engagement and fits different cultural contexts.	Overreliance on AI could reduce human interaction.	Enhance self-regulated learning, which allows teachers to focus on higher-order tasks.	High
[19]	Extended reality integration for performance-based experiential learning.	Ethical concerns in data usage for personalization.	Supports student development in immersive environments.	Moderate
[20]	Streamline tasks like grading, allowing teachers to focus on teaching.	Risks of inaccurate context/information in automated processes.	Requires human oversight for balanced experiences.	Low

Table I summarizes the role of AI in delivering personalized and adaptive learning experiences in higher education. It highlights how tools provide immediate feedback and customized content based on student performance, boosting engagement and comprehension. However, it also notes challenges such as the risk of reduced teacher-student interaction due to automation and ethical concerns in data usage and emphasizes the need for human oversight to maintain effective learning experiences.

Overall, the potential of AI in education is mainly adaptive personalization and time-saving efficiency. Adaptive tools customize instruction to diverse learner needs, interests, and

cultural contexts, helping close achievement gaps [18]. Systems that track emotional and cognitive engagement would be able to provide feedback to keep students focused and invested [19].

B. Assessment and Feedback Automation

Using automation in assessment enhances efficiency and scalability in diverse learning environments by transforming how teachers assign grades and provide feedback [21]. AI grading systems can deliver objective results compared to human grading, which often shows bias [21]. This also allows for a much faster and more efficient system, especially for students dissatisfied with traditional grading methods and the amount of time needed for results and feedback [21]. These systems enable timely personalized feedback that encourages self-learning and deeper engagement with the material [22, 23]. However, automated feedback struggles with writing tasks or assessments that don't align with rigid linguistic criteria, as they may fail to recognize diverse or unique ways students demonstrate understanding [24]. Carefully designed systems with adaptive features can overcome these limitations to significantly improve learning outcomes [22]. In legal education, hybrid intelligence systems provide formative high-level feedback while preserving the teacher's role in deeper instruction [23].

For automated feedback to succeed, it must account for varied student performances and learning styles, especially in tasks like writing that resist fixed syntactic patterns [24]. Students are more likely to accept AI feedback when they have prior experience with it and perceive it as fair and unbiased compared to human grading, particularly when outcomes are unfavorable [21]. Transparent system design and student-centered implementation are essential for broader adoption [25]. Generative AI systems like ChatGPT advance feedback by offering structured personalized guidance aligned with learning goals [25]. When paired with clear prompts and ethical boundaries, these tools support student autonomy without replacing the teacher's evaluative role [22, 25]. Over-reliance on AI risks undermining this balance, but thoughtful integration allows automation to handle routine tasks while freeing teachers to focus on complex, creative, and motivational aspects of education [23, 24].

Table II outlines the impact of AI-driven assessment and feedback systems, which focus on their ability to reduce grading bias and enhance efficiency through objective and timely feedback. It discusses how these systems support self-learning but struggle with non-rigid tasks like writing, which may not align with fixed criteria. The table shows the importance of adaptive and transparent designs to improve learning outcomes while preserving the teachers' role in deeper instruction.

TABLE II. AUTOMATED ASSESSMENT AND FEEDBACK: SUMMARY AND GRADE ASSESSMENT

No.	Key Points	Challenges	Effect on Teachers and Students	GRADE
[21]	AI gives objective feedback, reducing bias and time, which promotes self-learning	Students accept AI more if perceived as fair, but dissatisfaction with traditional methods still happens.	Faster feedback encourages deeper engagement, which is essential for scalability in large classes.	Moderate
[22]	Personalized feedback helps with learning goals and systems in legal education.	Struggles with non-rigid tasks like writing.	Improves outcomes with adaptive features, which frees teachers for more creative aspects.	Low
[23]	Formative feedback preserves the teacher's role in deeper instruction.	Over-reliance risks compromising balance.	Supports autonomy, which requires thoughtful integration.	High
[24]	Handles different student writing styles with NLP exploration.	Fails to recognize unique demonstrations of understanding.	Enhance learning experience for student-centered designs.	High
[25]	Generative AI like ChatGPT offers structured guidance with ethical boundaries.	Transparency issues in system design.	Wider adoption via trust and clear prompts.	Moderate

C. AI Academic Advising and Learning Analytics

AI-powered academic advising and learning analytics are revolutionizing higher education by providing personalized, data-driven support to enhance student outcomes. Machine learning models can achieve high accuracy in predicting course outcomes, enabling students to select courses that align with their academic strengths, which helps performance and reduces dropout rates [29]. Also, many AI chatbots offer 24/7 accessible guidance that allows for comprehensive responses to general academic and career queries, which helps promote education for diverse learners and those who have limited access to traditional advising [28]. Systems like the Artificial Intelligence Tutoring System (AITS) can use web technologies and decision trees to analyze student performance data, help identify struggling students, predict final grades, and recommend personalized learning paths for students [30].

These AI systems enhance student engagement and retention by offering scalable and personalized recommendations. For example, AI advising platforms analyze large datasets using academic records and career preferences to provide personalized courses and career pathway suggestions to help students with academic confidence and in making decisions [27]. Additionally, chatbots are also another option that can help streamline tasks like course registration and question handling, which would allow teachers to have more time and meaningful mentoring roles [26]. Studies highlight that factors like ease-of-use and trust significantly affect when adopting the use of these technologies, especially values like efficiency and accessibility [26]. AI can help make decisions and create synergies for students to use AI for guidance and help teachers with deeper and personalized support [28].

However, ethical challenges and data privacy concerns are still big problems. For example, AI advising systems rely on huge amounts of student data which would cause issues regarding compliance with regulations and bias results if there is not enough data of different types of students [27]. While personalized results are a key benefit, there is research that shows it may be less impactful and less efficient during the adoption stage, as the collection of data requires refinement

and time to become trustworthy [26]. These reasons and challenges show AI's potential in academic advising and learning analytics, but priority regarding data governance and bias is needed to ensure fair, effective, and efficient implementation across educational fields.

TABLE III. AI IN ACADEMIC ADVISING AND LEARNING ANALYTICS: SUMMARY AND GRADE ASSESSMENT

No.	Key Points	Challenges	Effect on Teachers and Students	GRADE
[26]	Chatbots streamline tasks, which helps boost adoption through ease of use.	Initial data collection requires clarification for trustworthiness.	Free teachers for mentoring and help with efficiency for diverse learners.	Moderate
[27]	Analyze the dataset for personalized learning career suggestions to improve confidence.	Risk of bias and data privacy if the dataset lacks diversity	Enhances retention and prioritizes data governance.	High
[28]	24/7 guidance via chatbots promotes accessibility.	Compliance with regulations is needed.	Empowers students with limited access to advising.	Low
[29]	High accuracy in predicting outcomes for course selection.	Less impactful during early adoption stages.	Reduces dropout rates and helps with strengths and benefits.	Moderate
[30]	Decision trees for performance analysis and recommendations.	Ethical challenges in large data reliance.	Identifies struggling students early and offers personalized paths.	Moderate

Table III shows the benefits of AI in academic advising and learning analytics, such as predicting course outcomes and providing 24/7 guidance through chatbots to enhance student retention and accessibility. It also addresses ethical challenges, which include data privacy risks and potential biases in datasets. This highlights the need for robust data governance to ensure fair and effective implementation across diverse educational settings.

D. Generative AI and the Challenge of Academic Integrity

Generative AI's use in higher education offers many possibilities but also causes serious challenges to academic integrity. Tools like ChatGPT support writing, research, and personalized learning yet increase risks of academic dishonesty through plagiarism and automated essay generation [32]. Students sometimes use these tools to skip genuine learning, producing assignments with minimal effort, while traditional plagiarism detection systems often fail to catch AI-generated content [33, 35]. Teachers worry about students submitting work lacking originality or understanding, especially since some AI can create entire assignments independently [33]. Ethical knowledge gaps and outdated institutional policies worsen these issues, underscoring the need for clear governance and updated integrity standards [31].

AI itself can also help maintain academic honesty. AI-based systems detect plagiarism, verify originality and provide feedback to promote ethical behavior which often outperforms older methods [32]. Teacher uses these tools to spot cheating patterns and confirm work authenticity [32]. However, relying heavily on automated systems raises concerns about false positives, reduced student agency and lack of trust due to privacy or fairness issues when tools are opaque or unreliable [33, 34].

To tackle these challenges, institutions must rethink assessment models to emphasize project-based learning, critical analysis and real-world applications, which are harder for AI to replicate and encourage deeper engagement [31, 35]. Yet implementing these changes globally faces hurdles, especially in resource-limited settings with inadequate infrastructure, insufficient training and resistance from students and faculty [32, 34]. Promoting responsible AI use demands more than policy updates; it requires fostering an academic culture that values integrity, boosts digital literacy and aligns assessments with both technological realities and educational goals [33].

TABLE IV. GENERATIVE AI AND ACADEMIC INTEGRITY: SUMMARY AND GRADE ASSESSMENT

No.	Key Points	Challenges	Effect on Teachers and Students	GRADE
[31]	Supports in writing research and plagiarism detection better than traditional methods.	Outdated policies and lack of ethical knowledge	Rethink assessments for critical analysis and help foster integrity	Moderate
[32]	Verifies originality and promotes ethical behavior.	Risks of dishonesty via automated generation through giving false positives	Use AI for detection but training in resource limited settings	High
[33]	Gives feedback that help with spotting cheating pattern.	Students skip learning but create lack of trust in more traditional tools	Emphasize project-based learning and boost digital literacy	Moderate
[34]	Maintain honesty through advanced systems.	Privacy/fairness issues which reduces student autonomy	Align assessments with tech realities and drive policy updates	Moderate
[35]	Shifts to holistic assessments are harder for AI to replicate	Problem in global implementation in under resource areas	Encourages deeper engagement and creates responsible AI use	Moderate

Table IV shows the dual role of generative AI in higher education which helps in supporting writing and research, while posing risks to academic integrity through plagiarism and automated content generation. It notes that AI-based plagiarism detection tools outperform traditional methods, but face issues like false positives and reduced student trust. The table advocates project-based assessments and updated policies to foster ethical AI use and deeper student engagement.

E. AI and Inclusion in Higher Education

AI is increasingly becoming more used in higher education institutions and showing more inclusivity and new ways to support learners with diverse needs. For example, robotic

assistants have been developed for children with ADHD which shows how AI can help users maintain attention and consistent engagement that could benefit students with cognitive challenges [36]. As well, smart technologies developed for older adults show the value of adaptive interfaces and principles where both are important for making accessible digital environments for people varying in physical and cognitive abilities [37]. These innovations highlight the need for participatory design approaches that prioritize user input from the start [36-37].

However, access to AI alone does not guarantee inclusion, as students are still facing barriers due to problems like digital literacy [39]. Studies show that individuals with visual impairments are more likely to engage with digital tools when they've had formal education and training in assistive technologies, which shows the need for targeted support in higher education [39]. AI tools can help with reading on smart screens and real-time feedback systems, which can save a lot of time. However, their effectiveness on user competence and the extent where institutional are ready to provide correct training are still questioned [39, 40]. Also, individual factors such as trust, technostress, self-efficacy, and personality traits can also influence whether students and teachers use AI tools at all [40].

AI-generated text tools are now common in academic settings, which raises concerns about fairness, transparency, and potential misuse [38]. Biased algorithms and decision-making processes risk having educational inequalities unless addressed through transparent design and clear data representation [38, 40]. Inclusive integration and admission of AI use require not only technical safeguards but also a cultural and constant commitment to accessibility and ethical practice [37, 38]. Overall, AI should be used as a tool to help all learners and not just the digitally privileged.

TABLE V. AI FOR INCLUSIVE EDUCATION: SUMMARY AND GRADE ASSESSMENT

No.	Key Points	Challenges	Effects on Teachers and Students	GRADE
[36]	Robotics design for ADHD children to maintain attention	Access doesn't guarantee inclusion, especially due to barriers of low digital literacy	Targeted support for cognitive challenges	Moderate
[37]	Adaptive interfaces for older adults and smart tech for varying abilities	Biased algorithms risk inequalities	Cultural commitment to accessibility can affect ethical practice	Moderate
[38]	AI text tools are common in academics, and real-time feedback saves time	Fairness/transparency concerns and potential misuse arise	Address biases through transparent design	High
[39]	Training aids for visually impaired people and engagement via education	Institutional readiness for training due to personal factors like trust	Bridge the digital divide and formal assistive tech education	Moderate
[40]	Influences adoption via self-efficacy/personality and saves time	Technostress and competence issues	Inclusive integration for all learners and not just the privileged	Moderate

Table V shows AI's potential to enhance inclusivity in higher education, such as through robotics for ADHD students and adaptive interfaces for diverse learners. It identifies barriers like low digital literacy and institutional readiness, which hinder equitable access. This table emphasizes the need for transparent design and targeted training to address biases and ensure AI supports all learners, not just the digitally privileged.

F. AI in Teaching Assistants and Virtual Classrooms

AI is changing virtual classrooms, allowing dynamic, interactive, and customized learning through AI teaching assistants. Incorporating AI into video material via tools can allow for real-time and context-aware interaction in class, which changes students from passive viewers to active participants [41]. These systems help explain concepts, visualization, and answer students' questions. As well, intelligent agents in an e-learning environment act as virtual tutors by providing feedback and encouraging interactive discussion [42]. These types of features allow for a successful mimicking of an aspect of human instruction and action [44]. AI teaching assistants also increase accessibility in education. Chatbots can act as a real-time response and learning companion, offering academic assistance 24/7 through a natural language interface [43]. Further studies show that conversational AI can reduce teachers' workload while supporting students' learning [42]. Robot-assisted systems in VR settings allow students to learn and practice language in a low-pressure and contextual environment, which encourages students' learning independence [45], which extends the AI role beyond the traditional classroom [43].

These systems are very personalized. For example, SAM feedback is based on an individual chat history and video context, which allows a tailored assessment; the system also provides personalized summaries for revision [41]. AI tools offer adaptive support across learning phases, such as generating quizzes and providing analytics to help guide teachers [44]. Also, AI isn't a replacement but an assisting tool that allows teachers to focus more on higher-order teaching, while AI handles routine tasks [42]. However, AI integration still poses challenges, such as students often using AI chatbot for basic tasks like formatting or rephrasing rather than learning these themself [43]. AI can also struggle with complex communication due to program limitations [45]. These issues show the need for AI that supports critical thinking and reason-based interaction [41]. Further development of AI should focus on AI as an education tool that complements human teachers and encourages active learning rather than passive dependence [44].

Table VI shows how AI teaches assistants and virtual classroom tools such as chatbots and VR systems that create interactive and personalized learning environments. It highlights their ability to reduce teacher workload and provide 24/7 support, but notes challenges like over-reliance on AI for basic tasks and limitations in complex communication. The table stresses the importance of AI as a complementary tool to promote active learning and critical thinking.

TABLE VI. AI TEACHING ASSISTANTS AND VIRTUAL CLASSROOMS: SUMMARY AND GRADE ASSESSMENT.

No.	Key Points	Challenges	Effects on Teachers and Students	GRADE
[41]	Tools to allow real-time interaction in videos and personalized summaries	Struggles with complex communication and risk of dependence	Turns passive viewers into active and personalized assessments	Moderate
[42]	Intelligent agents as tutors reduce workload.	Use of AI for basic tasks may hinder skill learning	Supports interactive discussion and enhances human teaching	High
[43]	Chatbots as 24/7 learning companions; natural language interface.	Limitations in program capabilities	Academic assistance at any time encourages independence	Low
[44]	Adaptive support with analytics and mimics human instruction.	The need for critical thinking focus.	Frees teachers for higher order tasks and allows for active learning	Moderate
[45]	Robot-assisted VR for low-pressure practice and learning.	Complex interactions challenging	Extends beyond classrooms and helps via contextual learning	Moderate

G. Institutional and Faculty Readiness

Institutional readiness for AI integration in education and healthcare depends on infrastructure, policy alignment and leadership support. Across sectors, institutions often lack the digital maturity required for large-scale AI adoption. For example, smart hospitals may often have a lot of investments for future technological development, but many public authorities and staff lack the expertise and training to use them effectively [46]. As well, educational institutions face challenges related to policy clarity, equitable access and resource allocation which are essential for the implementation of AI in higher education settings [48]. There are also infrastructure limitations such as specialized devices and connectivity which remain a significant barrier due to the large amounts of resources needed, even when administrative backing exists [49]. Without proper planning, enough resource allocation and community involvement, AI tools risk being underused or inequitably deployed [49, 50].

Faculty readiness is very important but also often overlooked, as teachers and health professionals are open to the concept and idea of AI, but their positive interactions often depend on targeted training and professional development [47]. Studies show that interested teachers often have steep learning curves due to insufficient structured assistance and ongoing training [47, 48]. Professional skills values such as autonomy, ethics and trust also influence people's uptake of AI tools. If these values were overlooked, resistance ensure [48]. As well, healthcare professionals must also possess digital literacy in order to counsel patients and practice in telemedicine environments [46]. Despite this requirement there are still skill deficiencies [46].

Individual attitudes and pedagogical models often influence AI uptake in high levels of self-efficacy and quality digital knowledge among teachers as this increases the likelihood that they will generate beneficial AI tool adoption [49]. However, internal readiness must be supplemented with institutional support through constant policies, mentorship programs and provision of technological infrastructure [49]. Perceived usefulness and believability of AI tools also affect the adoption by the faculty. If the AI tools are seen as credible and simple to use, then their adoption is much more likely [50]. Without trust, even the most advanced tools may be accepted hesitantly or rejected. Overall, readiness is not only about access to tools, but also an issue of organizational capacity, personal confidence and shared values. Professional development must go beyond functionality to address moral issues, career identity and structures of long-term care [48, 50]. Organizations that invest in meaningful training to establish ethics guidelines and foster collective experimentation will benefit from the advantages of AI in an equitable and sustainable way.

TABLE VII. INSTITUTIONAL AND FACULTY READINESS FOR AI INTEGRATION: SUMMARY AND GRADE.

No.	Key Points	Challenges	Effects on Teachers and Students	GRADE
[46]	Infrastructure for AI in healthcare and education and technological acceptance in adoption	Lack of expertise/training and lack of skills	Invest in digital maturity and equitable implementation	Low
[47]	Targeted training improves well-being and reduces workload	Steep learning curves if lacked support	Professional development beyond functionality	Low
[48]	Policy clarity and resource allocation and values like ethics influence adoption acceptance	Resistance to changes if autonomy overlooked	Mentorship programs address moral issues	Moderate
[49]	High self-efficacy boosts adoption and allows community involvement	Infrastructure barriers like connectivity	Supplement with institutional support and shared values	High
[50]	Trust in credible tools and mediating effects on practices	Hesitant acceptance without perceived usefulness	Foster experimentation and sustainable integration	Moderate

Table VII shows the factors affecting institutional and faculty readiness for AI adoption, which includes infrastructure, training and trust. It notes that, while AI can reduce workload and enhance teaching, barriers like limited digital literacy, resource constraints and resistance due to ethical concerns persist. The table underscores the need for comprehensive training, clear policies and community involvement to ensure equitable and sustainable AI integration.

V. Discussion

This review demonstrates that Artificial Intelligence (AI) has rapidly evolved from a supplementary tool to a core enabler of personalized and adaptive learning in higher education. Across the seven key topics analyzed, AI tools such as intelligent tutoring systems, generative platforms and predictive analytics consistently help and enhance student engagement, inclusivity and academic outcomes by personalizing and adapting content made for individual learning and needs [16-19]. The studies also show the transformative capacity of AI to close achievement gaps and offer equitable opportunities for students from diverse backgrounds or with special needs [12-14, 36-37].

However, the studies also highlight persistent ethical and practical challenges. Automated assessment tools and generative AI, while streamlining grading and feedback, risk diminishing the authenticity of assessments and student cognitive development if used without adequate safeguards [20, 31-35]. Academic integrity emerges as a major theme, with plagiarism detection and originality verification systems only partially mitigating risks [32-34]. However, issues of data privacy, algorithmic bias and transparency remain consistent across academic advising, learning analytics and inclusion initiatives [27-28, 38-40]. These challenges underscore the need for robust frameworks to ensure responsible implementation of AI technologies in education.

Institutional and faculty readiness also plays a decisive role in AI's success. While many teachers are open to AI adoption, gaps in digital literacy, professional development and infrastructure limit the full realization of AI's benefits [46–50]. Students' AI skills and teachers' perceptions of these skills do not always align, which may result in misjudgments of assessment authenticity or unintended over-reliance on automated systems. Addressing these disparities through targeted training, transparent communication and participatory design approaches can improve trust and efficacy in AI-driven systems [37-40]. Overall, the findings point to a dual imperative: scaling AI innovations to broaden access while embedding ethical safeguards and human oversight to preserve educational integrity.

VI. Conclusion

This review shows that artificial intelligence is no longer just an optional tool in higher education but a force that is actively reshaping how students learn, how teachers teach, and how institutions operate. Across different areas—such as assessment, advising, inclusion, and academic integrity, AI offers clear benefits: faster feedback, more personalized learning, and great support for diverse learners. However, the same technologies also create new problems that cannot be ignored, including risks to ethical practices, unequal access, and a growing dependence on automated systems.

A key message that emerges is that AI alone will not improve education. Its value depends on how well it is guided by human judgment, institutional policy, and shared responsibility among educators, students, and administrators. The findings suggest that successful use of AI requires three conditions: thoughtful design of learning and assessment, investment in digital skills for both teachers and students, and clear policies that protect fairness, privacy, and academic values.

Rather than seeing AI as a replacement for human roles, this review supports the idea that AI should work as a partner—reducing repetitive tasks while allowing teachers to focus on creativity, mentoring, and deeper learning. The next step for higher education is not to decide whether to use AI, but how to use it responsibly, so that technology enhances learning without weakening integrity or human connection.

Acknowledgment

The authors gratefully acknowledge the valuable support and cooperation of the advisors, academics, researchers, students, and staff who contributed to the completion of this research. The authors also wish to thank the Thailand Science Research and Innovation Fund and the supporting organizations for their assistance throughout the study.

References

[1] N. Duarte, Y. Montoya Pérez, A. J. Beltran and M. Bolaño García, "Use of Artificial Intelligence in Education: A Systematic Review," May 2023, doi: 10.46254/sa04.20230169.

[2] F. A. Bah, "Artificial Intelligence in Education System," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 439–443, Nov. 2024, doi: 10.48175/ijarsct-22863.

[3] H. Boumediene and M. Bouakkaz, "Artificial Intelligence in Education," *Advances in educational technologies and instructional design book series*, pp. 407–438, Dec. 2024, doi: 10.4018/979-8-3693-7220-3.ch014.

[4] K. I. K. Gyonyoru and J. Katona, "Student Perceptions of AI-Enhanced Adaptive Learning Systems: A Pilot Survey," pp. 93–98, Oct. 2024, doi: 10.1109/cando-epe65072.2024.10772884.

[5] P. Dubey, "Personalized Learning and Pedagogy," *Advances in educational technologies and instructional design book series*, pp. 81–102, Oct. 2024, doi: 10.4018/979-8-3693-6130-6.ch004.

[6] I. Ogunsakin, "Unlocking The Potential of Artificial Intellignce," Nov. 2024, doi: 10.33830/ijrse.v6i2.1698.

[7] K. Albinali, N. A. Iahad, A. F. Yusof and A. Abdulla, "A Systematic Literature Review of the Artificial Intelligence Role in Transformative Academic Advising: A Study of AI Applications in Higher Education," *International journal of academic research in business & social sciences*, vol. 14, no. 12, Dec. 2024, doi: 10.6007/ijarbss/v14-i12/24100.

[8] M. D. H. Wirzal, N. A. H. M. Nordin, N. S. A. Halim and M. A. Bustam, "Generative AI in Science Education: A Learning Revolution or a Threat to Academic Integrity? A Bibliometric Analysis," *Jurnal Penelitian dan Pengkajian Ilmu Pendidikan*, vol. 8, no. 3, pp. 319–351, Oct. 2024, doi: 10.36312/e-saintika.v8i3.2127.

[9] J. K. Wiredu, N. S. Abuba and H. T. R. Zakaria, "Impact of Generative AI in Academic Integrity and Learning Outcomes: A Case Study in the Upper East Region," *Asian Journal of Research in Computer Science*, vol. 17, no. 8, pp. 70–88, Jul. 2024, doi: 10.9734/ajrcos/2024/v17i7491.

[10] J. K. Wiredu, N. S. Abuba and H. T. R. Zakaria, "Impact of Generative AI in Academic Integrity and Learning Outcomes: A Case Study in the Upper East Region," *Asian Journal of Research in Computer Science*, vol. 17, no. 8, pp. 70–88, Jul. 2024, doi: 10.9734/ajrcos/2024/v17i7491.

[11] N. J. Francis, S. Jones and D. P. Smith, "Generative AI in Higher Education: Balancing Innovation and Integrity," *British Journal of Biomedical Science*, vol. 81, Jan. 2025, doi: 10.3389/bjbs.2024.14048.

[12] A. Hamdan, "The double-edged sword of AI-integrated education: an investigation into personalized and inclusive learning in higher education," pp. 381–391, Dec. 2024, doi: 10.4324/9781003606642-91.

[13] G. K. Ohri, "Advancing educational equity: the role of generative ai in creating accessible and inclusive higher education," *International journal of computer engineering and technology*, vol. 15, no. 6, pp. 1806–1815, Dec. 2024, doi: 10.34218/ijcet_15_06_154.

[14] A. P. Singh, R. Saxena and S. Saxena, "The Future of Learning: AI-Driven Personalized Education," *Asian journal of current research*, vol. 9, no. 4, pp. 207–226, Dec. 2024, doi: 10.56557/ajocr/2024/v9i49018.

[15] M. Abdullah, G. Rao, F. Sowell, V. J. Nirmal and S. Deb, "Optimizing Virtual Learning: Advanced Recommendations for an AI Teaching Assistant," Aug. 2024, doi: 10.18260/1-2--47822.

[16] A. K. Bawaneh, "AI Shaping the Future of Education: Science and Math Teachers' Satisfaction Level and Motivating Factors towards Integrating Artificial Intelligence in Teaching and Learning," *International Journal of Information and Education Technology*, vol. 15, no. 3, pp. 496–509, Jan. 2025, doi: 10.18178/ijiet.2025.15.3.2261.

[17] E. Elbasi, M. Nadeem, Y. I. Alzoubi, A. E. Topcu and G. Varghese, "Machine learning in Education: Innovations, impacts and ethical considerations," *IEEE Access*, vol. 13, pp. 128741–128770, Jan. 2025, doi: 10.1109/access.2025.3590134.

[18] K. S. McCarthy *et al.*, "Toward more effective and equitable learning: Identifying barriers and solutions for the future of online education.," *Technology Mind and Behavior*, vol. 3, no. 1, Mar. 2022, doi: 10.1037/tmb0000063.

[19] M. Hmoud, W. Daher and A. Ayyoub, "From experience to engagement: a mixed methods exploration of learning environments using artificial intelligence and extended reality," *Frontiers in Education*, vol. 10, Jul. 2025, doi: 10.3389/feduc.2025.1617132..

[20] A. Küçükuncular and A. Ertugan, "Teaching in the AI Era: sustainable digital education through ethical integration and teacher empowerment," *Sustainability*, vol. 17, no. 16, p. 7405, Aug. 2025, doi: 10.3390/su17167405.

[21] S. M. Jones-Jang, M. Chung, J. Choi, N. Kim and S. Lee, "Fairness perceptions of AI in grading systems: Examining how discontent with the status quo and outcome favorability reduce AI reluctance," *Computers and Education Artificial Intelligence*, p. 100419, May 2025, doi: 10.1016/j.caeai.2025.100419.

[22] D.-K. Lee and I. Joe, "A GPT-based Code Review System with Accurate Feedback for Programming Education," *IEEE Access*, p. 1, Jan. 2025, doi: 10.1109/access.2025.3581139.

[23] F. Weber, T. Wambsganss and M. Söllner, "Enhancing legal writing skills: The impact of formative feedback in a hybrid intelligence learning environment," *British Journal of Educational Technology*, Oct. 2024, doi: 10.1111/bjet.13529.

[24] M. Goldshtein, A. G. Alhashim and R. D. Roscoe, "An NLP-Based Exploration of Variance in Student Writing and Syntax: Implications for Automated Writing Evaluation," *Computers*, vol. 13, no. 7, p. 160, Jun. 2024, doi: 10.3390/computers13070160.

[25] G. Huesca, M. E. Elizondo-García, R. Aguayo-González, C. H. Aguayo-Hernández, T. González-Buenrostro and Y. A. Verdugo-Jasso, "Evaluating the potential of generative artificial intelligence to innovate feedback processes," *Education Sciences*, vol. 15, no. 4, p. 505, Apr. 2025, doi: 10.3390/educsci15040505.

[26] A. Saihi, M. Ben-Daya, M. Hariga and R. As'ad, "A structural equation modeling analysis of generative AI chatbots adoption among students and educators in higher education," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100274, 2024, doi: 10.1016/j.caeai.2024.100274.

[27] "AI-driven academic advising in higher education: Leveraging intelligent systems to personalize student support, improve retention and optimize career pathways," *CRSSS*, vol. 3, no. 2, pp. 229–248, Apr. 2025, doi: 10.59075/vy3v7k17.

[28] D. Akiba and M. C. Fraboni, "AI-supported academic advising: Exploring ChatGPT's current state and future potential toward student empowerment," *Educ. Sci.*, vol. 13, no. 9, p. 885, 2023, doi: 10.3390/educsci13090885.

[29] B. Shannaq, "The role of AI in university course registration in the Middle East: AI and machine learning approaches to improve academic performance," in *Proc. 2nd Int. Conf. Comput. Data Analyt. (ICCDA)*, Shinas, Oman, 2024, pp. 1–6, doi: 10.1109/ICCDA64887.2024.10867316.

[30] F. Grivokostopoulou, I. Perikos and I. Hatzilygeroudis, "Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance," in *Proc. IEEE Int. Conf. Teach. Assess. Learn. Eng. (TALE)*, Wellington, New Zealand, 2014, pp. 488–494, doi: 10.1109/TALE.2014.7062571.

[31] B. G. Acosta-Enriquez *et al.*, "AI in academia: How do social influence, self-efficacy and integrity influence researchers' use of AI models?," *Social Sciences & Humanities Open*, vol. 11, p. 101274, Jan. 2025, doi: 10.1016/j.ssaho.2025.101274.

[32] E. A. Obura and P. I. Emoit, "Artificial intelligence in academic writing and research skills in Kenyan Universities: Opportunities and challenges," *Africa Education Review*, pp. 1–23, Mar. 2025, doi: 10.1080/18146627.2024.2440351.

[33] P. D. Yadav, V. Madhukar, A. Dhamija and R. Manrai, "Understanding behavioral intentions of UG and PG students in Tier II Indian cities toward AI-technology adoption in higher education: a mixed-methods approach using the SOR model," *Cogent Social Sciences*, vol. 11, no. 1, Mar. 2025, doi: 10.1080/23311886.2025.2476044.

[34] A. D. Hernández *et al.*, "Mapping the individual, social and biospheric impacts of Foundation Models," *2022 ACM Conference on Fairness, Accountability and Transparency*, vol. 7, pp. 776–796, Jun. 2024, doi: 10.1145/3630106.3658939.

[35] A. Ateeq, M. Alzoraiki, M. Milhem and R. A. Ateeq, "Artificial intelligence in education: implications for academic integrity and the shift toward holistic assessment," Frontiers in Education, vol. 9, Oct. 2024, doi: 10.3389/feduc.2024.1470979.

[36] J. Berrezueta-Guzman, I. Pau, M.-L. Martin-Ruiz and N. Maximo-Bocanegra, "Assessment of a robotic assistant for supporting homework activities of children with ADHD," *IEEE Access*, vol. 9, pp. 93450–93465, Jan. 2021, doi: 10.1109/access.2021.3093233.

[37] J. Zhou et al., "Grand challenges of smart technology for older adults," International Journal of Human-Computer Interaction, pp. 1–43, Feb. 2025, doi: 10.1080/10447318.2025.2457003.

[38] Y. A. Mohamed, A. H. Mohamed, A. Kannan, M. Bashir, M. a. E. Adiel and M. A. Elsadig, "Navigating the Ethical Terrain of AI-Generated Text Tools: A review," IEEE Access, p. 1, Jan. 2024, doi: 10.1109/access.2024.3521945.

[39] T. Phochai, P. Setthasuravich, A. Pukdeewut and S. Wetchakama, "Bridging the Digital Disability Divide: Determinants of Internet Use among Visually Impaired Individuals in Thailand," Disabilities, vol. 4, no. 3, pp. 696–723, Sep. 2024, doi: 10.3390/disabilities4030043.

[40] L. Symasek, T. Yeazitzis, K. Weger and B. Mesmer, "Recent developments in individual difference research to inform the adoption of AI technology," Systems, vol. 13, no. 3, p. 156, Feb. 2025, doi: 10.3390/systems13030156.

[41] A. Bodonhelyi, E. Thaqi, S. Özdel, E. Bozkir and E. Kasneci, From Passive Watching to Active Learning: Empowering Proactive Participation in Digital Classrooms with AI Video Assistant. 2025, pp. 1–21. doi: 10.1145/3706598.3713513.

[42] K. Okanlawon, "From Automation to Personalization: A Comprehensive Survey on the Role of AI in Modern e-Learning", Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS), vol. 2, no. 1. Tech-Sphere Multidisciplinary International Journal (TSMIJ), pp. 1–17, Jul. 16, 2025. doi: 10.5281/zenodo.15960870.

[43] M. Usher and M. Amzalag, "From prompt to polished: Exploring Student–Chatbot interactions for academic writing assistance," *Education Sciences*, vol. 15, no. 3, p. 329, Mar. 2025, doi: 10.3390/educsci15030329.

[44] P. Kwan, R. Kadel, T. D. Memon and S. S. Hashmi, "Reimagining flipped learning via Bloom's taxonomy and Student–Teacher–GenAI interactions," *Education Sciences*, vol. 15, no. 4, p. 465, Apr. 2025, doi: 10.3390/educsci15040465.

[45] Y.-L. Chen, C.-C. Hsu, C.-Y. Lin and H.-H. Hsu, "Robot-Assisted Language Learning: Integrating Artificial Intelligence and Virtual Reality into English Tour Guide Practice," Education Sciences, vol. 12, no. 7, p. 437, Jun. 2022, doi: 10.3390/educsci12070437.

[46] A. I. Stoumpos, F. Kitsios and M. A. Talias, "Digital transformation in healthcare: technology acceptance and its applications," International Journal of Environmental Research and Public Health, vol. 20, no. 4, p. 3407, Feb. 2023, doi: 10.3390/ijerph20043407.

[47] N. Kshetri, "The Future of Education: Generative Artificial Intelligence's Collaborative Role With Teachers," *IT Professional*, vol. 25, no. 6, pp. 8–12, Nov. 2023, doi: 10.1109/MITP.2023.3333070.

[48] K. Huszárová, D. Šmida, A. Kmet, Š. Karolcik, and E. Hájková, "AI in the Classroom: How Teachers Perceive the Future of Education," in *2024 International Conference on Emerging eLearning Technologies and Applications (ICETA)*, Oct. 2024, pp. 206–211. doi: 10.1109/ICETA63795.2024.10850792.

[49] M. Li and E. Manzari, "AI utilization in primary mathematics education: a case study from a southwestern Chinese city," Education and Information Technologies, Jan. 2025, doi: 10.1007/s10639-025-13315-z.

[50] K. D. V. Prasad and T. De, "Generative AI as a catalyst for HRM practices: mediating effects of trust," Humanities and Social Sciences Communications, vol. 11, no. 1, Oct. 2024, doi: 10.1057/s41599-024-03842-4.

[51] B. Gundogan, N. Dowlut, S. Rajmohan, R. Mimi, C. Borrelli, M. C. Millip, C. Iosifidis, Z. Yagazie, U. Ginimol, M. Mathew, J. Alexander, R. Fowler and A. Agha, "Assessing the compliance of systematic review research works published in leading dermatology journals with the PRISMA statement guidelines: A systematic review," 2020, doi: 10.1016/J.JDIN.2020.07.007.

[52] S. Hullumani & R. Raghumahanti, "PRISMA Guidelines and Search Strategy," 2024, doi: 10.6084/m9.figshare.25013264.

[53] G. H. Guyatt, A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, et al., "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations," BMJ, vol. 336, no. 7650, pp. 924–926, 2008. doi: 10.1136/bmj.39489.470347.AD

[54] G. H. Guyatt, A. D. Oxman, G. E. Vist, R. Kunz, R. Falck-Ytter, Y. Alonso-Coello, et al., "GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables," J. Clin. Epidemiol., vol. 64, no. 4, pp. 383–394, 2011. doi: 10.1016/j.jclinepi.2010.04.026

A Secure Integrated Cloud Storage Framework Using Lorenz Chaotic Key Generation, AES-256-GCM, and PBFT-Based Blockchain Verification

Walde Rajesh Baliram[1*], Bashir Alam[2], Mohammad Najmud Doja[3]
Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India[1, 2]
Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India (Retired)[3]

***Abstract*—Cloud storage security remains a major challenge owing to threats related to confidentiality, integrity, and unauthorized access. In this study, a novel secure cloud storage system is proposed by integrating Lorenz 3D chaotic key generation, AES-256-GCM authenticated encryption, and blockchain verification based on PBFT. High-entropy keys were generated using the Lorenz chaotic system and then passed through SHA-256 to create secure 256-bit AES keys. The data is encrypted using the AES-256-GCM algorithm and stored in the cloud, and integrity metadata is stored in a blockchain that offers tamper detection, auditability, and Byzantine fault tolerance. The experimental results demonstrate strong cryptographic performance with near-ideal entropy (7.99), a high avalanche effect (50.05%), successful NIST randomness validation, and low blockchain overhead. The results of the security analysis prove resistance against brute force, statistical, differential, replay, known plaintext, chosen ciphertext, and Byzantine attacks. The proposed framework offers a secure, efficient, and reliable multilayered security solution for cloud storage systems.**

***Keywords*—Cloud security; AES-256-GCM; Lorenz chaotic system; blockchain; PBFT; data integrity; secure cloud storage; cryptography**

I. Introduction

Cloud computing has revolutionized the modern era of data management in terms of scalability, flexibility, and cost efficiency [1]. Cloud computing infrastructure is the foundation of the new economy, where companies rely on it to store sensitive data, such as financial records, healthcare information, and proprietary research data [2]. But when it comes to outsourcing data to external cloud service providers, there are significant security concerns, such as unauthorized access, data breaches, exposure of sensitive data, and privacy invasion [3].

The Advanced Encryption Standard (AES) is another symmetric encryption algorithm frequently employed to reduce the potential threat to confidentiality, as it offers high security assurance and a low computational load [4]. AES-256, especially, offers resistance to known cryptanalytic attacks, as well as brute force attacks under normal security conditions [5]. However, for centralized cloud storage systems, data encryption does not provide any data integrity checking or insider interference protection [6]. In addition, traditional cloud architecture has drawbacks, such as a lack of transparency, insufficient decentralization, and single points of failure, impacting trust in cloud-based systems [7].

The generation of keys is important in cryptography and is a critical component of security. The strongest encryption is not strong enough if the keys are predictable or not sufficiently random. In chaos-based cryptography, the high-entropy keys are generated using nonlinear dynamic systems that exhibit extreme sensitivity to initial conditions and have strong diffusion properties [8]. Chaos-based key generation with traditional encryption schemes has been proven to be a more random and efficient method [9]. The Lorenz 3D chaotic system is a complex nonlinear system that is highly sensitive to initial conditions and exhibits randomness, making it suitable for cryptographic applications. This makes it significantly more secure than low-dimensional chaotic systems and highly compatible with the latest encryption algorithms, including AES-256.

It is viewed as a decentralized solution to overcome trust and integrity problems in distributed systems using blockchain technology. The mechanisms of consensus, cryptographic hash linkage, and permanent ledger systems that blockchain offers can be used to store and record transactions in a verifiable manner without the possibility of being tampered with [10].

Recent studies have discussed the use of blockchain and cloud computing for access control auditing, logging security, and data integrity [11]. To resolve the problems of scalability in blockchain, researchers have suggested hybrid architectures, integration of cloud storage, and validation layers based on blockchain. In these systems, encrypted documents are stored in the cloud, and cryptographic hash values and transaction metadata are recorded on-chain.

Although cloud security has advanced, some areas of research could still be pursued. Existing AES-based cloud services offer confidentiality, but not decentralized integrity verification. Cloud-based frameworks using blockchain tend to lack cryptographic randomness enhancement, but are nonetheless tamper-proof. Cloud frameworks based on blockchain are not only tamper-proof, but also have no improvement in this regard with respect to cryptographic randomness. Chaos-based encryption methods provide more unpredictability but are not generally used in combination with distributed integrity validation mechanisms. Many existing systems lack formalized threat modelling or comparative security-performance evaluation. Limited research has been

*Corresponding author

conducted on incorporating authenticated encryption, chaotic key generation, and blockchain with PBFT within a single architecture.

The following contributions were made to overcome these limitations:

- Design and development of a secure cloud storage system with multi-level security, application of chaos-based cryptography, and verification by blockchain.
- Design of a Lorenz 3D chaotic key generation mechanism strengthened by the SHA-256 hash function.
- Incorporation of AES-256-GCM authenticated encryption with confidentiality and integrity.
- Implementation of a dual-hash integrity verification system for ciphertext and plaintext checks.
- Adoption of the PBFT consensus mechanism for Byzantine fault tolerance in blockchains.
- Complete experimental evaluation, including NIST randomness tests, entropy, avalanche effect, throughput, and latency measurements.
- Comparative analysis of existing AES-based and Chaos-based cloud security models.

The novelty of this work is the unified integration of the following four concepts uniformly: chaos theory-based key generation, authenticated encryption, dual hash integrity checking, and PBFT-based blockchain consensus to ensure secure cloud storage.

The structure of the rest of this study is as follows: A literature review of cloud security, chaos-based cryptography, and blockchain-based integrity verification is provided in Section II. In Section III, the proposed architecture for secure cloud storage is introduced. The methodology and algorithms are described in Section IV. Details of the experimental environment and setup are given in Section V. The experimental results and performance analyses are presented in Section VI. The threat model is introduced in Section VII, and a detailed security analysis is presented in Section VIII. A formal security analysis of the proposed framework is presented in Section IX. The comparison of the proposed approach with existing studies is made in Section X. The results and implications are discussed in Section XI. Finally, the conclusions and future research directions are presented in Section XII and Section XIII, respectively.

II. Literature Review

Data security remains a significant concern after cloud computing has transformed computing and storage services in recent years. Despite the use of encryption to reduce the risks of confidentiality, preliminary studies have pointed out that there are still some shortcomings, such as multi-tenancy, data exposure, and even user control over external resources [3], [12], [13]. Insider threats, improper access control, and post-storage tampering cannot be mitigated through encryption [14].

Symmetric encryption, particularly AES, is a fundamental aspect of secure cloud storage; however, its shortcomings in key generation and integrity checks have led to interest in other options. Because of its efficiency and robust cryptographic guarantees, AES is still the most popular symmetric encryption algorithm in cloud security architectures [15].

AES-256 is designed to be secure with respect to known cryptanalytic attacks and brute force attacks known to the designer, based on the known security assumptions. The recent research has also applied AES in the storage and transmission of information in cloud storage systems to protect the confidentiality of information [16]. AES is confidential but not necessarily decentralized, ensuring integrity verification and protection against unauthorized modifications in cloud storage systems.

Most existing schemes are built on traditional key generation schemes, which are susceptible to being compromised by insufficient entropy sources. The chaos-based cryptography utilizes a nonlinear dynamical system to improve the randomness and unpredictability of the key generation process [17]. The chaotic maps, such as logistic maps, are very sensitive to initial conditions and have positive Lyapunov exponents, matching the cryptographic randomness conditions [8], [18]. Research has shown that keys generated by the chaos technique are robustly diffused and confused, making them suitable for use to supplement the symmetric encryption technique [19]. Recently, studies have been carried out on using chaos-based methods to enhance unpredictability and protection against brute force attacks on cryptographic systems [20]. However, few studies have considered both chaos-based key generation and integrity models for blockchain-based cloud systems.

Blockchain technology has been called a new technology with the potential to be used to protect the integrity of data and distributed trust. The novel idea of decentralized consensus and hash-linked blocks to prevent tampering was introduced by Nakamoto [10]. Additional studies have led to the development of blockchain models and consensus algorithms for securing decentralized applications.

Alghuried and Wenhua provided a comprehensive overview of blockchain security concerns and privacy implications [21], [22]. The immutability of blockchain, its consensus-driven validation, and the decentralized system structure make it appropriate for cloud system integrity (Zheng et al., 2017). Even with these enhancements, several blockchain-based systems add performance overheads from the consensus protocol or complicated cryptographic primitives.

Some studies have focused on how to use the blockchain and cloud storage to enhance auditability and tamper resistance [16]. To improve scalability, blockchain-based cloud systems store large-scale, encrypted data off-chain and maintain blockchain metadata on-chain within the blockchain [11], [23]. However, blockchain integration with cloud services still faces many challenges related to performance overhead, transaction latency, and security concerns [24]. This has made lightweight blockchain architectures and hybrid storage solutions interesting potential solutions. However, large amount of data cannot be directly stored on the blockchain network due to

scalability and delay latency issues, and storage issues [25]. For this reason, more scalable solutions have been developed, which are hybrid solutions storing encrypted data off-chain with integrity metadata on-chain [26].

The application of chaos theory, authenticated encryption, blockchain technology, and distributed consensus has been investigated in recent studies to enhance cloud security. Rahman et al. proposed a chaos-enhanced AES key generation mechanism, which showed enhanced key randomness and unpredictability in an IoT environment [9]. Alatawi proposed a novel approach to decentralized authentication and authorization using smart contracts that run on a blockchain in cloud security [27]. In addition, Yao et al. and Liu et al. proposed optimal PBFT consensus mechanisms to enhance the throughput, latency, and Byzantine fault tolerance in the consortium blockchain system [28], [29]. Although these advancements reflect a considerable leap forward in secure cloud storage architectures, most current solutions focus on encryption, access control, or blockchain verification individually.

The proposed framework is distinct because of the combination of Lorenz 3D chaotic key generation and AES-256-GCM authenticated encryption, PBFT-based blockchain verification, and dual-hash integrity validation in a single cloud security framework.

III. Proposed System Architecture

The proposed system architecture is divided into six interoperable layers, as shown in Fig. 1. The proposed architecture is a multi-layered security architecture that combines symmetric encryption (AES-256-GCM), Lorenz 3D chaotic key generation, cloud storage, blockchain-based integrity verification using PBFT consensus, and a smart contract. These layers serve different functions to ensure the confidentiality, integrity, authentication, and traceability of the data. This multilayered design ensures a clear separation of concerns and separates cryptographic functions from blockchain consensus and the application logic. This modularity allows systems to be more resilient in the event of a compromise in one layer and to maintain a secure flow of data within and between the components. The layers are given below.

A. Application Layer

The Application Layer is the main layer where users interact with the proposed secure cloud storage system. It handles user authentication, access control, upload/download of files from/to the blockchain, and blockchain authorization. When a file is uploaded, the layer generates a Lorenz 3D chaotic key and applies AES-256-GCM authenticated encryption to protect the confidentiality and integrity of the data. The ciphertext is encrypted and stored in the cloud, with file hashes and transaction metadata stored on the blockchain, which is maintained on the PBFT protocol for integrity checks and auditing purposes. During the retrieval phase, AES-GCM authentication tag verification and blockchain validation are performed prior to decryption, which is robust enough to thwart attempts at unauthorized access, insider attacks, replay attacks, and data tampering.

B. Key Generation Layer

The Key Generation Layer is developed to produce a safe and high-entropy key of 256 bits from the Lorenz 3D chaotic system to a cryptographic hashing function called SHA-256. This blended approach yields the unpredictability and cryptographic strength of the generated keys. A highly sensitive, nonlinear sequence is generated by the Lorenz 3D chaotic system and quantized to be fed into SHA-256 to generate a secure, uniformly distributed AES-256 key for encrypting the vector.

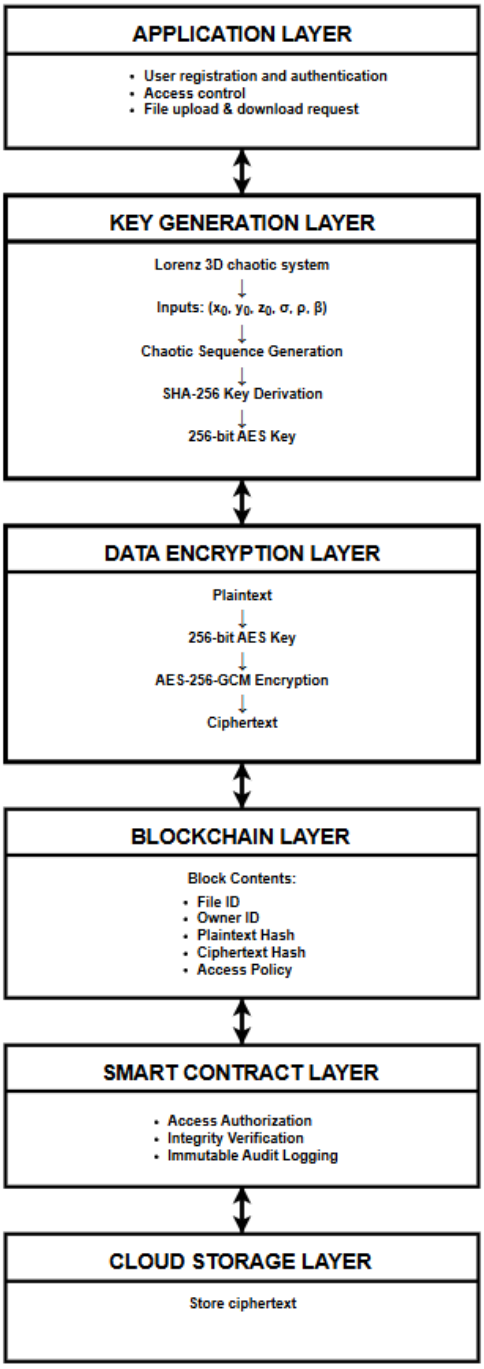


Fig. 1. System architecture diagram

C. Encryption Layer

The main security layer in the proposed framework that offers data confidentiality, integrity, and cryptographic randomness is the Encryption Layer. The chaotic Lorenz system is employed to produce a 256-bit chaotic key that is highly random, and then this 256-bit chaotic key is utilized for authenticated encryption with the AES-256-GCM. AES-GCM provides a ciphertext with a high degree of randomness and resistance to differential and statistical cryptanalysis. AES-GCM was created by AES to prevent modification or replay attacks on the ciphertext and ensure that the data remains intact. The integrity metadata and hash value of the encrypted ciphertext are recorded on the PBFT-based blockchain ledger for security checks and auditability, and the ciphertext is uploaded to the cloud server.

D. Cloud Storage Layer

Only encrypted data files are stored in the Cloud Storage Layer; no plaintext data is stored on the cloud servers.

E. Blockchain with PBFT Consensus Layer

Blockchain with the PBFT Consensus Layer provides auditability, integrity verification, and tamper-proof logging. The system does not store the entire file on-chain; instead, it stores the following:

- SHA-256 hash of the encrypted file
- SHA-256 hash of plaintext file
- Timestamp
- Previous block hash
- Block index

The blockchain layer offers security characteristics, such as immutability, tamper detection, audit trails, and decentralized verification. If any change is made to the stored file, the recalculated hash will not match the blockchain record, which will be the first sign of manipulation.

F. Smart Contract Layer

The Smart Contract Layer specifies and implements access control policies. Smart contracts, which are built on blockchains, offer complete control over access to cloud data. Downloading the resource is only possible after a successful on-chain authorization of a download request to prevent unauthorized access, even by privileged insiders of the cloud. Blockchains' smart contracts ensure data integrity, enable decentralized access control, store immutable hash, and deliver transparent auditability. Both serve crucial functions in enhancing the privacy and security of cloud storage.

IV. METHODOLOGY

A. System Workflow

The workflow of the proposed secure cloud storage system is defined by the entire cycle of data, from uploading to access and verification (Fig. 2). This system consists of client-side encryption, decentralized access control, blockchain-based integrity verification, and cloud storage to ensure confidentiality, integrity, and accountability.

Step 1: User Registration. Users and owners of the data are registered into the system with unique identifiers. Blockchain credentials are issued to each user to safely interact with smart contracts.

Step 2: AES key Generation. The AES key is generated based on a highly sensitive and nonlinear sequence generated by the Lorenz 3D chaotic system. Then, the SHA-256 hashing algorithm was used to create a 256-bit encryption key.

Step 3: File Encryption. The data owner selects a file to be stored in the cloud. The data is encrypted on the client using AES-256-GCM encryption, which ensures confidential data storage.

Step 4: Secure File Upload. The encrypted file is then uploaded to the cloud storage. The access permissions of the files are defined by the data owner before they are uploaded.

Step 5: Metadata Storage in Blockchain. A cryptographic hash of the plaintext file and a cryptographic hash of the encrypted file are generated using a hashing algorithm called SHA-256 to help verify the integrity of the file. The file hash and file ownership information were recorded on the blockchain. The access policy and file identifier, along with a hash, are stored and verified on the blockchain via a smart contract.

Step 6: Access Request by User. Once the user requests access to a stored file, it is sent to the blockchain layer. The smart contract then receives the access control policy and proceeds to determine whether the requesting user has access to that file.

Step 7: Decentralized Access Control Decision. Access is granted or denied based on smart contract verification. The information requested, and the decision made is logged on to the blockchain as an audit trail which is auditable and non-repudiable.

Step 8: Secure File Download. Once authenticated, encrypted files are downloaded from the cloud storage and sent to the authorized end user. The cloud service provider does not know the content of the files and does not control access to the files.

Step 9: File Tamper Detection. It computes the SHA-256 hash of the downloaded encrypted file prior to decryption. This hash is compared with the original hash that was added to the blockchain. If any discrepancy was found, it meant that the data or the process were corrupted, and the process was killed.

Step 10: File Decryption. If tamper detection test is passed successfully, the encrypted file is decrypted at the client side using the AES key to reconstruct the plain text file.

Step 11: File Integrity Verification. SHA-256 computes the hash of the decrypted file in the system. This hash is compared with the hash of the original plaintext file stored in the blockchain. If there is any discrepancy, it means that there is some corruption or tampering with the data, and the process is killed.

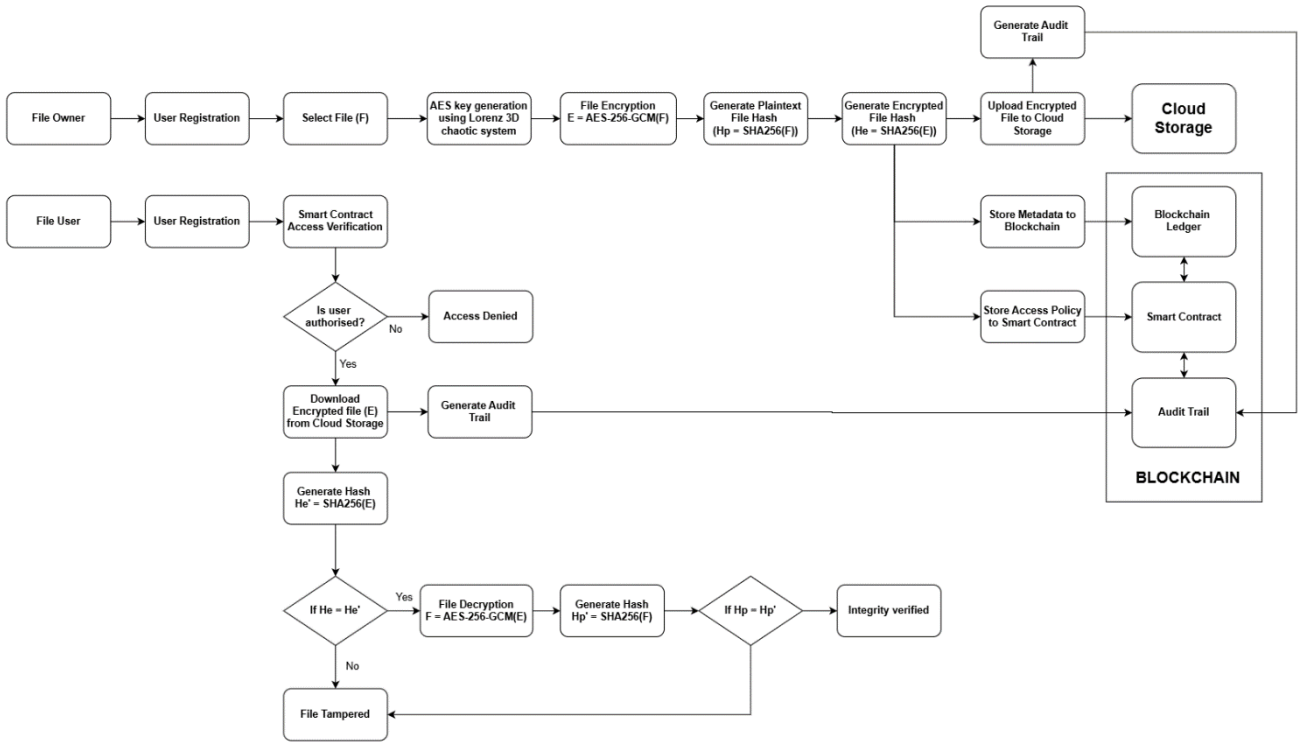


Fig. 2. System workflow diagram

B. Algorithms

To implement the proposed secure cloud storage framework, five algorithms are used to deliver confidentiality, integrity, authentication, and fault tolerance. Algorithm 1 produces a cryptographic key with high entropy by applying the Lorenz 3D chaotic system. Algorithm 2 is used to perform AES-256-GCM authenticated encryption, which provides data confidentiality and integrity. Algorithm 3 securely uploads encrypted files onto cloud storage and records integrity metadata on the blockchain. Algorithm 4 uses the Practical Byzantine Fault Tolerance (PBFT) consensus protocol for blockchain network validation and committing metadata transactions. Lastly, Algorithm 5 retrieves the encrypted files from the cloud storage and checks for integrity on both the encrypted and decrypted files based on blockchain-stored metadata. These algorithms create a robust, transparent, and unalterable cloud storage system. For clarity and ease of understanding, the notations and abbreviations used in the proposed algorithms, along with their descriptions, are provided in Table IX at the end of the study.

Algorithm 1: Key Generation Using Lorenz 3D Chaotic System

The purpose of this algorithm is to generate an AES key using the Lorenz 3D chaotic system [30].

Input:

x0, y0, z0: Initial conditions

σ, ρ, β: Lorenz control parameters

h: Step size

N: Number of iterations

T: Number of transient samples to discard

L: Required key length (256 bits)

Output:

K: 256-bit chaotic key

Begin

1. x ← x0
2. y ← y0
3. z ← z0
4. S ← EmptySequence
5. for i ← 1 to N do
6. x ← x + h × σ × (y - x)
7. y ← y + h × (x × (ρ - z) - y)
8. z ← z + h × (x × y - β × z)
9. Append(S, x)
10. end for
11. S ← RemoveFirstTElements(S, T)
12. B ← EmptyBitstream
13. for each value s in S do
14. x_norm ← |s| mod 1
15. if x_norm > 0.5 then
16. bit ← 1
17. else
18. bit ← 0
19. end if

```
20.     Append(B, bit)
21.     if Length(B) ≥ L then
22.         break
23.     end if
24. end for
25. B ← FirstLBits(B)
26. K ← ConvertBitsToBytes(B)
27. return K
End
```

Algorithm 2: AES-256-GCM Authenticated Encryption

The purpose of this algorithm is to securely encrypt user data using the AES-256-GCM authenticated encryption scheme specified in the Advanced Encryption Standard (AES) [4] and Galois/Counter Mode (GCM) [31].

```
Input:
  P: Plaintext divided into 128-bit blocks {P1, P2, ..., Pm}
  K_AES: 256-bit AES encryption key
  IV: 96-bit initialization vector or nonce
  AAD: Associated authenticated data
Output:
  C: Ciphertext
  T: 128-bit authentication tag
Begin
1. RoundKeys ← KeyExpansion(K_AES)
2. J0 ← IV || 0^31 || 1
3. C ← EmptyCiphertext
4. for i ← 1 to m do
5.     Ji ← inc32(Ji-1)
6.     Si ← AES_Encrypt(K_AES, Ji)
7.     Ci ← Pi ⊕ Si
8.     Append(C, Ci)
9. end for
10. H ← AES_Encrypt(K_AES, 0^128)
11. Y ← AAD || C || len(AAD) || len(C)
12. Divide Y into 128-bit blocks {Y1,Y2,…, Yn}
13. X0 ← 0^128
14. for j ← 1 to n do
15.     Xj ← (Xj-1 ⊕ Yj) · H over GF(2^128)
16. end for
17. G ← Xn
18. T ← AES_Encrypt(K_AES, J0) ⊕ G
19. return (C, T)
End
```

Algorithm 3: Secure File Upload

The purpose of this algorithm is to encrypt user data safely with the help of a chaos-based AES-256 key, save the ciphertext in the cloud, and store metadata about integrity in the blockchain for tamper-proof verification.

```
Input:
  F: Plaintext file
  UserCred: User authentication credentials
  x0, y0, z0: Initial conditions of Lorenz
3D chaotic system
  AAD: Associated authenticated data
Output:
  UploadStatus
Begin
1. if Authenticate(UserCred) = FALSE then
2.     return "Access Denied"
3. end if
4. K_chaos ← Lorenz3D_KeyGeneration(x0,y0,z0)
5. K_AES ← SHA256(K_chaos)
6. IV ← CSPRNG_Generate_Unique_Nonce(96 bits)
7. EncParams ← Encrypt({x0,y0,z0}, MasterKey)
8. StoreKeyMetadata(FileID, EncParams)
9. (C,T) ← AES256_GCM_Encrypt(F,K_AES,IV, AAD)
10. H_plain ← SHA256(F)
11. H_cipher ← SHA256(C)
12. StoreCloud(C, T, IV)
13. Metadata ← {
      FileID,
      H_plain,
      H_cipher,
      Timestamp
    }
14. Tx ← CreateBlockchainTransaction(Metadata)
15. ExecuteSmartContract(Tx)
16. UploadStatus ← StoreBlockchainRecord(Tx)
17. return UploadStatus
End
```

1) Secure chaotic parameter management: The initial conditions (x_0, y_0, z_0) of the Lorenz system are considered as confidential cryptographic parameters. These parameters are not stored in plain text in the cloud storage layer or blockchain ledger to prevent disclosure. Rather, they are encrypted with a master key specific to the user and stored in a secure key-management module. An encrypted parameter record is linked

with the appropriate file identifier during file upload. If the user is authenticated successfully during the retrieval of the file, the encrypted parameters are recovered and decrypted securely in local mode, which helps to regenerate the chaotic key sequence and derive the AES-256 encryption key. As a result, the confidentiality of the chaotic parameters is maintained, and the key is regenerated reliably by authorized users.

2) Nonce generation and management: A unique 96-bit nonce (IV) is generated for each file upload using a cryptographically secure random number generator (CSPRNG) to prevent nonce reuse in AES-256-GCM. This unique nonce is tied only to the encryption session in which it is used, and is not repeated when using the same AES key. The nonce is saved along with the ciphertext and authentication tag in the cloud storage and recalled at the decryption phase. The probability of nonce collision is negligible due to the following reasons: new AES key is generated each time the uploaded file is uploaded, and fresh nonce is used every time the data is encrypted. With this nonce management strategy, AES-GCM assures confidentiality and authentication while mitigating on nonce reuse.

Algorithm 4: PBFT-Based Blockchain Metadata Verification

The purpose of this algorithm is to validate and commit blockchain transactions with file integrity metadata using Practical Byzantine Fault Tolerance. The algorithm verifies that only valid ciphertext hashes, plaintext hashes, timestamps, and audit records are accepted after the validators in the consortium reach an agreement.

Input:
  Tx: File metadata transaction
  H_cipher: Ciphertext hash
  H_plain: Plaintext hash
  T: Timestamp
  FileID: File identifier
  V: Validator node set {v1, v2, ..., vn}
Output:
  B_i: Committed blockchain block
  ConsensusStatus: Accepted or Rejected
Begin
1. Tx ← {
    FileID,
    H_cipher,
    H_plain,
    T,
    UserID
  }
2. SignTransaction(Tx)
3. PrimaryNode ← SelectPrimary(V)
4. if VerifyTransaction(Tx) = FALSE then
5. return Rejected
6. end if
7. B_i ← Hash(B_(i-1) || Tx || T)
8. Broadcast(PRE_PREPARE, v, s, Digest(Tx))
9. for each validator v_i ∈ V do
10. if ValidatePrimaryNode() = TRUE
    and VerifyDigest(Tx) = TRUE
    and VerifySequenceNumber(s) = TRUE
    and VrifyPreviousBlockHash() = TRUE
    and VerifyTransaction(Tx) = TRUE then
11. Broadcast(PREPARE,v,s,Digest(Tx),NodeID)
12. end if
13. end for
14. if PrepareCount ≥ 2f then
15. Broadcast(COMMIT,v,s,Digest(Tx),NodeID)
16. end if
17. if CommitCount ≥ 2f + 1 then
18. Blockchain ← Blockchain ∪ B_i
19. ConsensusStatus ← Accepted
20. else
21. ConsensusStatus ← Rejected
22. RaiseAlert("Consensus Failure")
23. end if
24. return (B_i, ConsensusStatus)
End

Algorithm 5: File Retrieval and Verification

The purpose of this algorithm is to extract encrypted data from cloud storage and check the integrity of these data before and after decryption using the blockchain metadata.

Input:
  FileID: Unique file identifier
  UserCred: User authentication credentials
  K_AES: AES-256 decryption key
  AAD: Associated authenticated data
Output:
  F: Plaintext file
  RetrievalStatus: Verified or Rejected
Begin
1. if Authenticate(UserCred) = FALSE then

```
2.      return "Access Denied"
3. end if
4. EncParams ← RetrieveKeyMetadata(FileID)
5. (x0,y0,z0) ← Decrypt(EncParams, MasterKey)
6. K_chaos ← Lorenz3D_KeyGeneration(x0,y0,z0)
7. K_AES ← SHA256(K_chaos)
8. (C, T, IV) ← RetrieveCloud(FileID)
9. H_plain_ch,H_cipher_ch) ← QueryBlockchain(FileID)
10. H_cipher_local ← SHA256(C)
11. if H_cipher_local ≠ H_cipher_ch then
12.    RaiseAlert("Ciphertext Tampering Detected")
13.    RetrievalStatus ← Rejected
14.   Abort
15. end if
16. F ← AES256_GCM_Decrypt(C,K_AES,IV,AAD,T)
17. if DecryptionStatus = AuthenticationFailed then
18.     RaiseAlert("AES-GCM Authentication Failed")
19.     RetrievalStatus ← Rejected
20.     Abort
21. end if
22. H_plain_local ← SHA256(F)
23. if H_plain_local ≠ H_plain_chain then
24.    RaiseAlert("Plaintext Integrity Violation")
25.    RetrievalStatus ← Rejected
26.     Abort
27. end if
28. RetrievalStatus ← Verified
29. return (F, RetrievalStatus)
End
```

C. Secure Chaotic Parameter Management

The initial conditions of the Lorenz system (x_0, y_0, z_0) are treated as confidential cryptographic parameters. To prevent disclosure, these parameters are not stored in plaintext in the cloud storage layer or blockchain ledger. Instead, they are encrypted using a user-specific master key and stored within a secure key-management module. During file upload, the encrypted parameter record is associated with the corresponding file identifier. Upon successful user authentication during file retrieval, the encrypted parameters are securely recovered and decrypted locally to regenerate the chaotic key sequence and to derive the AES encryption key. Consequently, the confidentiality of the chaotic parameters is preserved while ensuring reliable key regeneration for the authorized users.

V. Experimental Setup

The following configuration was used to build a prototype system of the hybrid cloud storage solution and evaluate it in a sound-testing environment.

A. Software Environment

The software stack was selected for its cryptographic efficiency and ability to handle complex chaotic-map computations with high precision.

- Programming Language: Python 3.x was used because of its flexibility and extensive support for numerical and cryptographic libraries.
- Cryptographic Library: PyCryptodome was employed specifically for its AES-256-GCM implementation, which provides data confidentiality and authentication features.
- Hash Function: SHA-256 (via the hashlib library) was used to generate tamper-proof file digests and to distil the Lorenz chaotic state variables into fixed-length 256-bit keys.
- Blockchain Model: To enable secure logging of file metadata and verification hashes, a custom implementation of a blockchain using Python was developed for prototyping purposes. This implementation aims to demonstrate the PBFT consensus mechanism for verifying integrity in a controlled environment.
- Blockchain Consensus Method: Practical Byzantine Fault Tolerance was added to ensure that transactions are final and that the ledger is consistent across all the distributed nodes.
- Operating System: Windows 11.
- Integrated Development Environment: PyCharm 2025.2.1.1.

B. Hardware Configuration

The following hardware was used to obtain consistent performance in the execution of the chaotic integration and encryption algorithms:

- Processor: Intel(R) Core i7 processor
- RAM: 16 GB
- Storage: SSD-based storage to reduce the latency of I/O operations for large datasets.

C. PBFT Implementation

The proposed architecture is based on a custom consortium blockchain implementation with Python for experimentation to demonstrate integrity verification using PBFT in a secure cloud storage architecture. Implementation attempts to analyze the feasibility and performance influence of integrating the PBFT consensus into AES-256-GCM encryption and Lorenz chaotic key generation.

The experiment was conducted in a controlled local distributed environment with one client node, one primary (leader) validator node, and three replica validator nodes, totaling four PBFT validator nodes. The PBFT protocol is executed in the usual way: first, a pre-prepare phase, followed by a prepare phase, and finally a commit phase, to ensure that blockchain transactions that include ciphertext hashes, integrity metadata, timestamps, and audit records are validated. A transaction is committed if it is approved by $(2f + 1)$ validator nodes, with at most $(f \leq \frac{n-1}{3})$ Byzantine nodes.

This consortium blockchain design assumes that the validator nodes are maintained by various participating organizations or entities, including cloud administrators, organizational clients, and audit authorities, rather than a single centralized cloud provider. This distributed trust model can achieve Byzantine fault tolerance and prevent malicious insiders from unilaterally modifying the integrity of the metadata.

The locally distributed implementation was intentionally designed as a lightweight prototype to test the latency of the blockchain, the validation of transactions, the feasibility of the consensus, and the performance of the integration with the proposed encryption framework. Therefore, the reported blockchain latency values are experimental and proof-of-concept results, conducted on a small scale in a controlled test environment, and not in a large-scale deployment with a cloud-based environment.

Currently, the proposed framework is not designed to ensure the large-scale scalability of the proposed PBFT protocol. In the cloud-scale scenario, the number of validator nodes could increase the network overhead of the practical communication complexity of PBFT as $(O(n^2))$. However, for a typical consortium-based enterprise cloud environment, the number of trusted validator nodes is relatively small, and in practical circumstances, integrity verification and auditability applications are suitable for the PBFT.

D. Test Dataset

The framework was tested using a multi-format dataset to assess its performance at different levels of data redundancy and correlation. Experiments were carried out with file sizes ranging from 10 to 50 MB, such as:

- Standard text files
- High-correlation image datasets
- Unstructured binary files
- Complex PDF documents

All tests were conducted under controlled laboratory conditions to obtain consistent results. The main points of the experimental design were reproducibility and complete security validation. Each experiment was performed several times, and the average of the performance metrics was calculated to obtain the final performance metrics to maintain the statistical validity of the experiments and reduce the influence of random errors or outlier values.

VI. Experimental Results

A. Performance Evaluation of Metrics

The metrics used to evaluate the framework in this section are encryption, decryption, upload, download, blockchain consensus latency, and overall throughput.

1) Measurement methodology: The experimental tests were conducted on a locally distributed setup with file sizes ranging from 10 to 50 MB, and without using external networks to avoid the influence of network variations. All time-related measurements were reported in seconds (s).

a) Latency Calculation (T_{total}): The total processing time for a single file transaction is defined as the sum of its independent computational stages:

$$T_{total} = T_{encryption} + T_{upload} + T_{blockchain}$$

where,

- $T_{encryption}$: is the duration required for the Lorenz 3D chaotic key generation, SHA-256 distillation, and the AES-256-GCM encryption process.
- T_{upload}: is the time taken to transfer the 128-bit blocks of ciphertext to the hybrid cloud storage module.
- $T_{blockchain}$: is the time required for the PBFT consensus protocol to validate the metadata and commit the SHA-256 integrity hashes to the distributed ledger.

b) Throughput Calculation ($Throughput$): Throughput is defined as the number of file transactions completed per unit of time and is calculated as:

$$Throughput = \frac{Number\ of\ Files}{Total\ Time}$$

where, $N = 1$ representing a single file of a specific size processed from initialization to blockchain commitment.

2) Performance analysis: A detailed performance analysis of the proposed system was performed and executed for various file sizes ranging from 10 to 50MB. The experimental results for the different performance metrics are summarized in Table I and presented graphically in Fig. 3.

TABLE I. Performance Evaluation of Metrics

Metric	10MB	20MB	30MB	40MB	50MB
Chaotic Latency (Sec)	0.07	0.08	0.07	0.08	0.07
Encryption Latency (Sec)	0.05	0.08	0.13	0.18	0.23
Upload Latency (Sec)	0.08	0.12	0.18	0.22	0.27
Blockchain Latency (Sec)	0.01	0.01	0.01	0.01	0.01
Total Latency (Sec)	0.20	0.28	0.39	0.49	0.58
Throughput (Files/Sec)	4.95	3.59	2.59	2.05	1.74

As shown in Table I and Fig. 3, the proposed framework can maintain its performance when the file size increases from 10 to 50 MB. The total latency increases gradually from 0.20 sec to 0.58 sec due to higher encryption and upload times for larger files. By contrast, chaotic key generation latency is almost stable (0.07- 0.08 sec), and the blockchain latency is constant (0.01 sec), showing low overhead of the PBFT-based

verification process. As the file size increased, the throughput decreased from 4.95 files/s to 1.74 files/s, as anticipated, because larger files require more processing and transmission time.

Fig. 3. Performance evaluation of metrics chart.

The performance of the proposed system is compared with that of existing methods in terms of latency, throughput, and security overhead. In order to make a fair comparison, maximum file size used in the performance evaluation was 50MB. The comparative results are presented in Table II and Fig. 4.

TABLE II. PERFORMANCE METRICS AND CUMULATIVE SECURITY OVERHEADS.

Metric	Cloud system	Cloud + AES system	Cloud + AES + Chaotic system	Proposed system
Latency (Sec)	0.27	0.50	0.57	0.58
Throughput (Files/Sec)	3.73	2.01	1.76	1.74
AES Overhead (%)	0	85.73	85.73	85.73
Chaotic Overhead (%)	0	0	14.29	14.29
Blockchain Overhead (%)	0	0	0	1.20

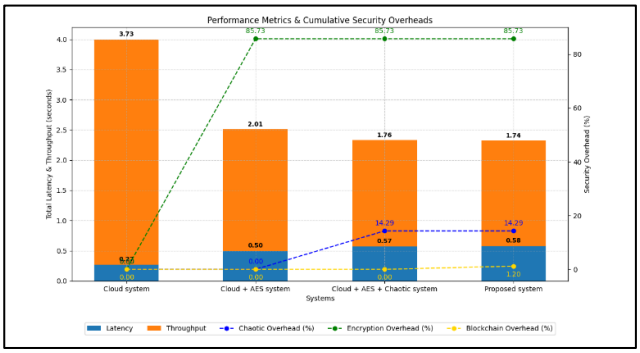


Fig. 4. Performance metrics and cumulative security overheads.

As shown in Table II and Fig. 4, the proposed system causes only a minor latency increase from 0.57 sec to 0.58 sec when compared with the Cloud + AES + Chaotic system. The slight increase is due to the extra verification layer added to the blockchain, which is based on the PBFT protocol. There is also a slight drop in throughput from 1.76 files/s to 1.74 files/s, demonstrating that blockchain integration has a negligible effect on the processing efficiency. Although the proposed system introduces AES, chaotic key generation, and blockchain verification overheads, the proposed blockchain overhead is only 1.20%, which shows that the proposed system provides increased integrity, auditability, and Byzantine fault tolerance at a small performance cost.

B. Chaotic Key Randomness Analysis

The Lorenz 3D chaotic key sequences were evaluated using several randomness evaluation measures, including Shannon entropy, bit balance analysis, chi-square testing, autocorrelation analysis, and key sensitivity analysis, as listed in Table III.

TABLE III. STATISTICAL ANALYSIS OF GENERATED CHAOTIC KEY SEQUENCE.

Category	Metric / Test	Result	Interpretation
Entropy Analysis	Shannon Entropy	7.99	Near-ideal randomness close to the theoretical maximum for Shannon entropy H=8
Bit Balance Analysis	Number of 0s	128242	Nearly balanced binary distribution
	Number of 1s	127758	Indicates low statistical bias
Chi-Square Analysis	Chi-Square Value	229.50	Near-uniform byte distribution with minimal deviation
Autocorrelation Analysis	Lag 0	1.00000	Perfect self-correlation at zero lag
	Lag 1	0.00245	Near-zero correlation
	Lag 2	0.00101	Near-zero correlation
	Lag 3	-0.00212	Near-zero correlation
	Lag 4	0.00128	Near-zero correlation
	Lag 5	0.00181	Near-zero correlation
	Lag 6	-0.00148	Near-zero correlation
	Lag 7	-0.00133	Near-zero correlation
	Lag 8	-0.00227	Near-zero correlation
	Lag 9	-0.00165	Near-zero correlation
Sensitivity Analysis	Key Sensitivity	0.51562	Strong chaotic sensitivity and diffusion behavior
NIST SP 800-22 Tests	Frequency Test	0.33877 (PASS)	Balanced distribution of bits
	Block Frequency Test	0.80179 (PASS)	Uniformity across bit blocks
	Runs Test	0.21520 (PASS)	Absence of abnormal repetitive patterns
	Longest Run Test	0.75065 (PASS)	Random distribution of continuous bit runs
	Approximate Entropy Test	0.52712 (PASS)	High sequence complexity and unpredictability

The results shown in Table III demonstrate that the Lorenz 3D chaotic key sequence exhibits good randomness and cryptographic features. The entropy value of 7.99 is close to the theoretical maximum of 8, and the nearly balanced distribution of 0s and 1s and the low chi-square value indicate that there is no statistical bias. In addition, the near-zero autocorrelation values and high key sensitivity indicate high

unpredictability and dependence on the initial conditions. The NIST SP 800-22 tests were passed successfully, thereby demonstrating that the keys produced were statistically random and appropriate for secure cryptographic use.

C. Ciphertext Statistical Analysis

A ciphertext statistical analysis is conducted to assess the effectiveness of the proposed encryption framework in creating statistically random encrypted data. To analyze the encryption diffusion and statistical security, the ciphertexts were tested by measuring the entropy, chi-square, autocorrelation, and avalanche effects (Table IV).

TABLE IV. STATISTICAL EVALUATION OF CIPHERTEXT

Category	Metric / Test	Result	Interpretation
Entropy Analysis	Shannon Entropy	7.99	Near-ideal ciphertext randomness close to the theoretical maximum H=8
Chi-Square Analysis	Chi-Square Value	295.75	Near-uniform ciphertext distribution with low statistical deviation
Autocorrelation Analysis	Lag 0	1.00000	Perfect self-correlation at zero lag
	Lag 1	0.00345	Near-zero correlation
	Lag 2	0.00104	Near-zero correlation
	Lag 3	0.00160	Near-zero correlation
	Lag 4	0.00188	Near-zero correlation
	Lag 5	0.00123	Near-zero correlation
	Lag 6	-0.00632	Near-zero correlation
	Lag 7	0.00194	Near-zero correlation
	Lag 8	-0.00840	Near-zero correlation
	Lag 9	-0.00400	Near-zero correlation
Avalanche Effect Analysis	1-bit Plaintext Change	50.05%	Strong diffusion and nonlinearity
	2-bit Plaintext Change	50.03%	Consistent ciphertext diffusion
	3-bit Plaintext Change	49.99%	Near-ideal avalanche behavior
	4-bit Plaintext Change	49.99%	Stable diffusion performance
Overall Ciphertext Behavior	Average Autocorrelation	Near Zero	Strong statistical independence between ciphertext bytes

The results presented in Table IV demonstrate that the proposed framework provides strong resistance to a wide range of security threats. The large key space (2^{256}), high entropy (7.99), and avalanche effect (50.05%) indicate robustness against brute-force, statistical, and differential attacks. In addition, AES-256-GCM authenticated encryption effectively protects against known-plaintext and chosen-ciphertext attacks, whereas dual-hash verification and blockchain immutability ensure reliable detection of data tampering and replay attacks.

D. Key Space Analysis and Histogram

The proposed scheme is based on the Lorenz 3D chaotic system, which has high sensitivity to generate chaotic sequences, and processes the chaotic sequence using the SHA-256 hash function to obtain a fixed-length AES-256-GCM key. The last encryption key is:

$$K = SHA256(f(x_0, y_0, z_0))$$

where, $f(x_0, y_0, z_0)$ is the output of the Lorenz chaotic system, and K is the key of AES (256 bits).

The chaotic system has a large parameter space and is very sensitive to the initial conditions; however, the key space of the final AES key specifies the space that can be used for cryptography.

Since SHA-256 outputs a 256-bit digest, the maximum effective key space is (2^{256}). This key space is too large and computationally impractical for exhaustive searches using existing classical computing methods. The key space histogram is shown in Fig. 5.

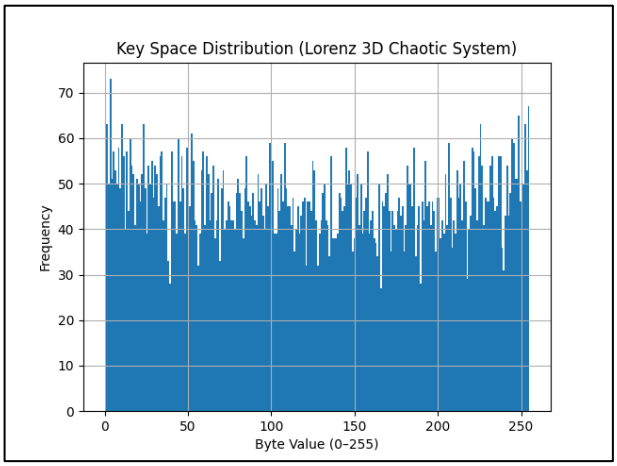


Fig. 5. Key space histogram

Fig. 5 illustrates that the proposed system achieves a key space of (2^{256}). The histogram exhibits a nearly uniform distribution of key space values evenly distributed throughout all 256 possible byte values, suggesting high randomness and no statistical bias in the key-space. This provides excellent resistance to both brute-force and statistical attacks.

E. Plaintext Histogram

The histogram of plaintext file is shown in Fig. 6.

In Fig. 6, the histogram shows a nonhomogeneous data distribution, has noticeable peaks, reflects inherent data patterns, and has a high degree of statistical redundancy. This irregularity makes the plaintext vulnerable to attacks, such as frequency analysis, statistical attacks, and pattern recognition.

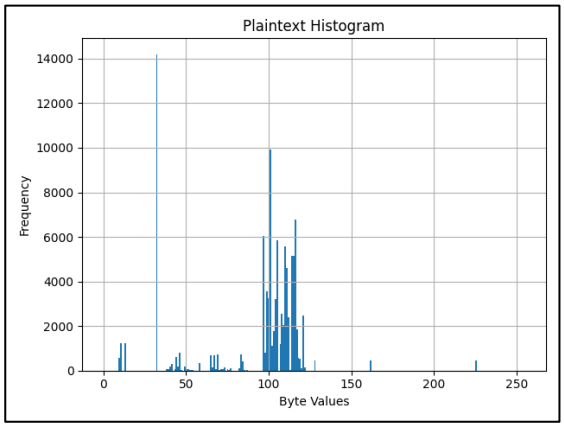


Fig. 6. Plaintext histogram

F. Ciphertext Histogram

The histogram of ciphertext generated by the proposed system is shown in Fig. 7.

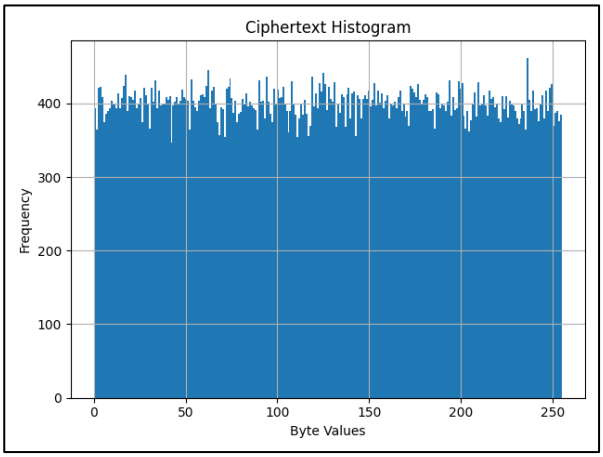


Fig. 7. Ciphertext histogram

The ciphertext histogram in Fig. 7 is almost unbiased, shows an absence of visible patterns, has a flattened frequency distribution curve, and features an equal distribution of all 256-byte counts. This indicates that diffusion, confusion, and statistical randomness were high. This chaotic key is used to increase the randomness of the AES cipher, thereby making the histogram more homogeneous. Thus, chaotic enhancement offers a considerable degree of uncertainty, randomness, and resistance to differential analyses.

G. Attacks and Resistance

Different multi-file real-attack simulations were performed, including unauthorized access, file tampering, blockchain tampering, brute-force, statistical, differential, known-plaintext, chosen ciphertext, replay, and Byzantine fault-tolerance attacks. The proposed framework exhibited excellent anti-attack capabilities against various attack types. Lorenz chaotic key generation, AES-256-GCM encryption and PBFT blockchain offer robustness against cryptographic, statistical, and system-level attacks. The security performance against various attacks is presented in Table V.

TABLE V. SECURITY PERFORMANCE AGAINST VARIOUS ATTACKS

Attack Type	Evaluation Metric	Result	Security Outcome
Unauthorized Access Attack	Access Control Success	100%	Access denied
Data Tampering Attack	Integrity Detection Rate	100%	Tampering detected
Blockchain Tampering	Block Validation Failure	100%	Invalid block rejected
Brute-Force Attack	Key Space	2^{256}	Computationally infeasible
Statistical Attack	Entropy	7.99 bits	Near-ideal randomness
Differential Attack	Avalanche Effect	50.05%	Strong diffusion
Known Plaintext Attack	Key Recovery Probability	Negligible	Resistant
Chosen Ciphertext Attack	Authentication Failure Detection Rate	100%	Attack Prevented
Replay Attack	Detection Rate	100%	Attack prevented
Man-in-the-Middle Attack	Hash Verification Success	100%	Attack detected
Byzantine Fault Tolerance	Maximum Faulty Nodes Tolerated	$f \leq \frac{n-1}{3}$	Consensus Maintained

Table V shows that the proposed framework offers robust resistance against several cryptographic and blockchain attacks. The user authentication and access control mechanisms prevent unauthorized access to the stored data. A large key space of (2^{256}), avalanche effect of 50.05% and entropy value of 7.99 assure strong resistance to brute force, statistical and differential attacks. The combination of Lorenz chaotic key generation and AES-256-GCM provides an additional layer of security against known plaintext and chosen ciphertext attacks. Blockchain Timestamp Validation, AES-GCM Authentication Tags, and Dual-Hash Integrity Verification help prevent replay attacks, data tampering, and ciphertext modification. Moreover, blockchain tampering is detected using immutable ledger records and hash verification mechanisms. Finally, PBFT ensures accurate network functioning even in situations where malicious validator nodes are present. Consensus is guaranteed as long as the number of faulty nodes does not exceed one-third of the total validators ($f \leq (n-1)/3$), demonstrating the robustness of the proposed framework against Byzantine failures. Overall, the results confirm that the proposed framework provides confidentiality, integrity, authenticity, and fault tolerance in cloud-storage environments.

VII. THREAT MODEL

The threat model defines the assumptions, attack surfaces, trust boundaries, and security goals of the proposed secure cloud-storage framework. The primary intent of this threat model is to evaluate how well the proposed framework can guarantee confidentiality, integrity, authenticity, auditability, and decentralized trust in realistic threat scenarios.

A. Adversary Assumptions

The adversary is assumed to have polynomial-time computational power and can attempt to attack various components of the system. An attacker can perform any of the following actions:

- Intercept communication channels
- Capture encrypted data packets
- Replay previously transmitted ciphertext
- Modify stored cloud data
- Attempt brute-force key recovery
- Perform statistical analysis on ciphertext
- Compromise a subset of blockchain validator nodes
- Attempt unauthorized access to cloud resources

This adversary is assumed to be capable of completely using the public communication network but lacking a legitimate cryptographic key or secret chaotic parameters.

B. Trust Assumptions

The following are the key assumptions of the proposed framework:

- AES-256-GCM remains computationally secure.
- SHA-256 is collision and pre-image-resistant.
- Initial chaotic parameters remain secret.
- Fewer than one-third of the PBFT validator nodes were malicious.
- It is assumed that the authentication mechanisms have been properly implemented.
- The random initialization vectors were generated securely.

These assumptions are based on widely used cryptographic and distributed-system security models.

C. Security Goals

The proposed framework is designed to achieve the following security goals (Table VI):

TABLE VI. SECURITY GOALS

Security Goal	Description
Confidentiality	Prevent unauthorized disclosure of plaintext
Integrity	Detect unauthorized modifications
Authentication	Verify legitimate users
Auditability	Maintain immutable access records
Availability	Ensure reliable data access
Non-Repudiation	Prevent denial of performed actions
Byzantine Fault Tolerance	Resist malicious validator behavior

The combination of cryptographic protection and decentralized verification in the framework provides multilayer security in a cloud storage environment.

VIII. SECURITY ANALYSIS

The security strength of the proposed framework is studied within the context of an extensive threat model of external attackers, insider attacks, and attempts to tamper with the data. The framework aims to maintain data confidentiality, integrity, authentication, and auditability through encryption using AES-256-GCM with Lorenz chaos and blockchain-assisted integrity verification (IV) techniques.

A. Overview of the Security Objectives

The proposed framework uses the Lorenz 3D chaotic system to generate a chaotic key, AES-256-GCM symmetric encryption, and a PBFT-based blockchain system to achieve multi-dimensional security. It has the following major security objectives:

- The data were confidential, as ensured through AES-256-GCM encryption.
- Assign a random value to the 'key' to achieve key unpredictability (using Lorenz chaotic dynamics).
- Integrity & Immutability: Ensured with blockchain hashing.
- The authentication & consensus are enforced through PBFT.
- Attack Resistance: Against brute-force, known-plaintext, chosen-plaintext, replay, statistical, Byzantine, and side-channel attacks, unauthorized access, file tampering, and blockchain tampering.

Let the system be defined as:

$$C = E_K(P), \quad K = H(f(x_0, y_0, z_0))$$

where, (x_0, y_0, z_0) are Lorenz's initial conditions and $H(\cdot)$ is SHA-256.

B. Unauthorized Access Resistance

Unauthorized access resistance is provided by authenticated encryption, blockchain access verification, and secure key management. The key used to encrypt the data is the AES-256-GCM encryption key, which is dynamically generated from the Lorenz chaotic system and is not stored as plain text in the cloud environment.

Only authenticated and authorized users with valid cryptographic credentials can decrypt the stored ciphertext successfully. Moreover, verification on the blockchain does not allow any unauthorized changes to be made in the access records or metadata of transactions to enhance protection against unauthorized attempts to access the cloud storage.

C. Ciphertext and File Tampering Resistance

The framework uses a dual integrity verification mechanism: AES-256-GCM authentication tag validation and blockchain-based SHA-256 ciphertext hash validation. If any part of the ciphertext, nonce, authentication tag, or content of the stored file changes, the authentication fails, or the hash of the blockchain does not match. In addition, the ciphertext hashes stored in the blockchain were verified before decryption. Therefore, when unauthorized changes are made, they are immediately discovered and rejected. This mechanism strengthens the resistance against:

- Ciphertext tampering
- Malicious modification
- Unauthorized data manipulation
- File integrity attacks

D. Blockchain Tampering Resistance

The PBFT-based blockchain layer offers distributed integrity checks and BFT. The number of malicious nodes that can be accommodated by a consensus by PBFT is limited to $(f \leq \frac{n-1}{3})$ validator nodes.

Blockchain metadata, such as the ciphers and transactions, is immutable and cannot be altered without the agreement of the validators that are part of the consensus. Therefore, the integrity records and transaction histories would be impossible for the attackers to change without compromising most of the validator nodes.

E. Brute-Force Resistance

The proposed framework uses AES-256-GCM authenticated encryption with a 256-bit chaotic key generated by the Lorenz 3D chaotic system. The total key space is (2^{256}), which is computationally infeasible to exhaustively search using current classical computing capabilities.

Moreover, the Lorenz chaotic system is highly sensitive to the initial conditions, leading to greater unpredictability. The slightest change in the chaotic parameters produces entirely different encryption keys, which makes the system even harder to break using brute-force and key prediction.

F. Statistical Attack Resistance

The generated ciphertext exhibits the following properties based on statistical analysis:

- Near-ideal Shannon entropy (H=7.99)
- Balanced statistical distribution
- Strong avalanche characteristics

In addition, the autocorrelation values of the ciphertext are close to zero for all nonzero lags, which further proves, on a statistical level of significance, that the ciphertext bytes are statistically independent. These results showed resistance to:

- Frequency analysis
- Statistical cryptanalysis
- Differential attacks
- Pattern recognition attacks

G. Differential Attack Resistance

The strong diffusion and confusion properties of the proposed encryption framework provide resistance to differential attacks. Experimental avalanche effect analysis showed that small changes in the plaintext resulted in enormous changes in the ciphertext, and the average value of the avalanche effect was nearly 50%, which is the theoretical ideal.

The Lorenz chaotic key generation mechanism further enhances the nonlinear diffusion characteristics and the unpredictability. In this way, the attacker will not be able to establish meaningful relationships between the differences in plaintext and ciphertext, making the system more resistant to differential cryptanalysis.

H. Known-Plaintext Attack Resistance

The proposed framework can prevent known plaintext attacks using dynamic chaotic key generation and AES-256-GCM authentication encryption. In the event that an adversary acquires partial plaintext/ciphertext pairs, the encryption key cannot be reconstructed because the Lorenz chaotic parameters remain concealed, the SHA-256 key derivation is one-way, and the AES-256-GCM delivers computational confidentiality.

Additionally, high ciphertext randomness and near-zero autocorrelation make it difficult for an attacker to detect exploitable statistical relationships between the known plaintext and ciphertext data.

I. Chosen-Ciphertext Attack Resistance

The framework includes authenticated encryption using AES-256-GCM and integrity verification using blockchain, which resists adaptively chosen ciphertext attacks. AES-GCM validates the authentication tag before decrypting data.

If the ciphertext, nonce, or authentication tag is modified in any way, decryption fails: $Dec_K(C') = \perp$

Hence, any malicious manipulation of the ciphertext does not provide useful information regarding plaintext data.

Furthermore, the hash of the ciphertexts stored on the blockchain is checked before the block is decrypted, allowing for early detection of unauthorized changes to the ciphertexts. Together, these mechanisms enhance resistance to IND-CCA2 adversaries and ciphertext forgery attacks.

J. Replay Attack Resistance

Replay attack resistance is implemented using unique nonces for AES-GCM and transaction validation in the blockchain. Every encryption operation results in different nonce and transaction identifiers, so that the same ciphertext cannot be reused. Blockchain metadata validation ensures the authenticity and integrity of the transaction before it is decrypted during retrieval from the file. Duplicated or outdated transactions are rejected, thus avoiding a replay attack.

In addition, replay protection has been improved with unique AES-GCM nonces and blockchain timestamps. As each file is encrypted with a different nonce and with a unique blockchain transaction record, replayed ciphertexts or duplicated upload requests will be identified via metadata verification/timestamp validation.

K. Byzantine Fault Tolerance Resistance

To overcome malicious or compromised validators, the proposed consortium blockchain uses a PBFT consensus protocol. Using PBFT, the blockchain can continue to function properly, even if some of the validator nodes are malicious or fail unexpectedly. A transaction is committed only when at least $(2f + 1)$ validator nodes approve the block.

This means that Byzantine attackers have no ability to tamper with the blockchain integrity record, audit logs, or access verification metadata without compromising more than $(2/3)$ of the validator nodes.

This mechanism is strengthened by:

- Distributed trust
- Consensus reliability
- Blockchain integrity
- Resistance against insider consensus manipulation

L. Insider Attack Resistance

The proposed framework helps to reduce insider threats by implementing the Encrypted cloud storage, Blockchain-based access verification, Immutable audit logging, and Authenticated encryption.

Cloud administrators and malicious insiders cannot access plaintext data without the corresponding chaotic-encryption keys. Moreover, every file action is logged as an immutable transaction on the blockchain, allowing for open and transparent auditing and accountability of information. Blockchain integrity verification and AES-GCM authentication can detect any unauthorized attempts to modify the data.

M. Overall Security Discussion

The Lorenz 3D chaotic key generation, AES-256-GCM authenticated encryption, SHA-256 integrity verification, and PBFT blockchain consensus contribute to the security of the proposed framework. The experimental tests yielded the following results:

- Near-ideal entropy
- Low statistical correlation
- Strong avalanche behavior
- Successful NIST randomness validation was achieved.

The results indicate that the proposed framework is resistant to brute-force, statistical cryptanalysis, replay, ciphertext tampering, insider, and manipulation attacks on the blockchain.

N. Comparative Security Analysis

The data were encrypted using AES-256-GCM to ensure their confidentiality. The SHA-256 hashes of the encrypted and plaintext files are stored in the blockchain, guaranteeing data integrity. Blockchain immutability ensures that access policies and audit logs cannot be changed, which, in turn, ensures non-repudiation and accountability of the data. AES encryption protects data privacy and keeps those who are not authorized to access the plaintext data, such as cloud providers. Smart contracts are decentralized access controls, data integrity with irrevocable hashes, and auditability on blockchains. A comparative security analysis of the proposed and conventional systems is presented in Table VII.

TABLE VII. COMPARATIVE SECURITY ANALYSIS

Feature	AES System	AES + Chaotic System	Proposed System
Confidentiality	Yes	Yes	Yes
Integrity Verification	Limited	Limited	Yes
Blockchain Auditability	No	No	Yes
Chaotic Randomness	No	Yes	Yes
PBFT Consensus	No	No	Yes
Replay Protection	No	Limited	Yes
Byzantine Fault Tolerance	No	No	Yes
Dual Hash Verification	No	No	Yes
NIST Validation	No	Partial	Yes

The comparative results presented in Table VII demonstrate that the proposed framework provides a more comprehensive security solution than existing approaches. Unlike existing schemes, which are mostly concerned with chaotic encryption and key space enhancement, the proposed scheme further includes PBFT consensus, Byzantine fault tolerance, replay protection, blockchain-based auditability, and dual-hash integrity check. Although maintaining comparable entropy (approx. 7.99) and a large key space (2^{256}), the proposed system offers stronger resistance to data tampering, replay attacks, and blockchain-related threats. The findings indicate that the combination of Lorenz chaotic key generation, AES-256-GCM encryption, and blockchain verification using PBFT can significantly improve the security and trust of cloud storage systems.

IX. FORMAL SECURITY ANALYSIS

A. Security Assumptions

The proposed framework relies on the following assumptions:

- AES-256-GCM is IND-CCA2 secure when nonces are unique.
- SHA-256 is collision-resistant and preimage-resistant.
- The Lorenz chaotic parameters remain secret and unpredictable, ideally derived from high-entropy sources or secure key exchange protocols during system initialization to prevent their compromise.
- PBFT consensus is secure if fewer than one-third of validator nodes are Byzantine.
- Validator nodes are distributed across independent entities in a permissioned consortium blockchain.

B. IND-CCA2 Security of the Encryption Layer

Theorem 1: The encryption layer of the proposed framework achieves IND-CCA2 security under the assumption that AES-256-GCM is a secure authenticated encryption scheme and that nonces are never reused with the same key.

Game 0: Real IND-CCA2 Game

The challenger generates the encryption key: $K = H(f(x_0, y_0, z_0))$

where, $f(x_0, y_0, z_0)$ denotes the Lorenz chaotic key.

The adversary A is given access to: $Enc_K(\cdot)$ and $Dec_K(\cdot)$

The adversary submits two equal-length plaintexts: (P_0, P_1)

The challenger selects: $b \in \{0, 1\}$

and returns: $C^* = Enc_K(P_b, N, AAD)$

The adversary may continue querying the decryption oracle, except that it may not query the exact challenge ciphertext C^*. Finally, the adversary outputs b'.

The adversarial advantage is as follows:

$$Adv_A^{IND-CCA2} = \left|\Pr[b' = b] - \frac{1}{2}\right|$$

Game 1: Replace Derived Key with Uniform Random Key

In this game, the derived key $K = H(f(x_0, y_0, z_0))$

is replaced with a uniformly random 256-bit key as follows:

$$K' \leftarrow \{0, 1\}^{256}$$

If the adversary distinguishes Game 0 from Game 1, then it can distinguish the derived key from a uniformly random AES key.

$$\text{Thus: } |\Pr[G_0] - \Pr[G_1]| \leq \epsilon_{KDF}$$

where, ϵ_{KDF} denotes the advantage of distinguishing the SHA-256-derived chaotic key from a uniformly random 256-bit key.

Game 2: AES-GCM IND-CCA2 Game

In Game 2, encryption and decryption are performed using AES-256-GCM with a uniformly random key K'.

AES-GCM provides authenticated encryption. Therefore, any modified ciphertext is rejected unless the adversary produces a valid authentication tag.

For a 128-bit authentication tag: $\Pr[Forge] \leq 2^{-128}$

Since AES-GCM is assumed IND-CCA2 secure under nonce uniqueness, the adversary cannot distinguish whether the challenge ciphertext encrypts P_0 or P_1 with non-negligible probability.

Therefore: $\Pr[G_2] = \frac{1}{2}$

and: $Adv_A^{IND-CCA2} \leq \epsilon_{KDF} + Adv_{AES-GCM}^{IND-CCA2}$

Since both terms are negligible: $Adv_A^{IND-CCA2} \approx 0$

Conclusion: The encryption layer is IND-CCA2 secure under the assumptions that AES-256-GCM is IND-CCA2 secure, nonces are unique, and the derived chaotic key is computationally indistinguishable from a random 256-bit key.

C. Blockchain Integrity Security

Theorem 2: The blockchain integrity layer detects ciphertext metadata tampering with an overwhelming probability, assuming that SHA-256 is a collision-resistant hash function.

The blockchain stores the ciphertext hash: $H_{\text{cipher}} = SHA256(C)$

During retrieval, the system computes: $H_{\text{local}} = SHA256(C')$

The file is accepted only if: $H_{\text{local}} = H_{\text{cipher}}$

If an adversary modifies the ciphertext: $C \rightarrow C'$

Then successful undetected tampering requires: $SHA256(C') = SHA256(C)$

where, $C' \neq C$

This implies a collision in SHA-256.

Therefore: $Pr[Undetected\ Tampering] \leq Adv_{SHA256}^{Collision}$

Since SHA-256 collision resistance is assumed secure: $Pr[Undetected\ Tampering] \approx 0$.

D. PBFT Consensus Security

Theorem 3: The PBFT-based blockchain maintains consensus safety if fewer than one-third of the validator nodes are Byzantine nodes.

For n validator nodes, PBFT tolerates: $f \leq \frac{n-1}{3}$ Byzantine nodes.

A block is committed only if at least: $2f + 1$ validators approve the transaction.

Therefore, an adversary can alter the blockchain metadata only if it controls more than one-third of the validator nodes or violates the PBFT consensus.

$$\text{Thus: } Pr[PBFT\ Failure] \leq Adv_{PBFT}$$

Under the stated validator trust model: $Adv_{PBFT} \approx 0$

E. Overall Security Bound

The overall framework security follows from the separate security properties:

$$Adv_{System} \leq Adv_{AES-GCM}^{IND-CCA2} + Adv_{SHA256}^{Collision} + Adv_{PBFT}$$

where,

- $Adv_{AES-GCM}^{IND-CCA2}$ represents the encryption-layer confidentiality advantage
- $Adv_{SHA256}^{Collision}$ represents the probability of undetected metadata tampering
- Adv_{PBFT} represents the probability of violating PBFT consensus assumptions

Unlike the previous version, blockchain integrity and PBFT consensus were not included as hybrid games in the IND-CCA2 proof. Instead, they are treated as separate system-level security properties.

Since each term is negligible under the stated assumptions: $Adv_{System} \approx 0$.

X. Comparative Study Table

Table VIII summarizes the main features of this study and some published studies and illustrates some important differences between them. Most current approaches are either encryption or blockchain-verification-driven, ignoring high-dimensional chaotic key generation and adequate formal security evaluation. The goal of the framework is to fill these key gaps by combining these elements and offering something that is more complete and provably secure than those proposed in the current literature.

The comparative results presented in Table VIII show that the proposed framework provides a more comprehensive and integrated security solution than existing approaches. The proposed framework integrates Lorenz 3D chaotic key generation, AES-256-GCM authenticated encryption, blockchain-based integrity verification, and PBFT consensus into a single architecture, whereas most previous studies focused on a single aspect of encryption, access control, or blockchain verification. This integration improves the privacy, integrity, audibility, and Byzantine fault tolerance of data while providing robust storage security in the cloud and efficient metadata verification. As a result, the proposed framework provides more security resilience and a comprehensive defense-in-depth protection for cloud storage systems.

TABLE VIII. Comparative Study Table

Features/ System	[11]	[32]	[33]	[34]	Proposed System
Architecture	Fully decentralized blockchain network	Blockchain + cloud server	Blockchain + cloud	Blockchain-enabled cloud architecture	Multi-layered architecture
Encryption Method	CP-ABE	SHA-256 + ECDSA	ABE	ChaCha20-Poly1305	AES-256-GCM
Key Generation Method	Bilinear mapping-based cryptography	--	ABE	ChaCha20-Poly1305	Lorenz chaotic 3D system
Blockchain Role	Access control + integrity	Transaction verification	Data verification	Access control + metadata	Decentralized integrity verification, immutable audit logging, and PBFT-based consensus for Byzantine fault tolerance.
Storage Model	Distributed cloud	Cloud server	Cloud + blockchain metadata	Cloud storage	Encrypted files are stored in cloud storage with integrity hashes and metadata maintained on a PBFT-based blockchain.
Performance Observation	Fully decentralized architecture increases processing overhead	Cryptographic verification increases security overhead	Blockchain verification improves integrity but adds latency	It surpasses AES-GCM and RSA in various performance metrics.	Detected evaluated attack scenarios, demonstrating strong security resilience, reliable integrity verification, and effective Byzantine fault tolerance
Key Features	Authorization, decentralized storage	Secure medical record sharing	Secure data sharing and verification	Improved throughput and lower latency	Attack detection and prevention, AES-256-GCM authenticated encryption, 10D hyperchaotic key generation, blockchain-based integrity verification, and PBFT-based Byzantine fault tolerance.

XI. Discussion

The experimental findings show that the proposed secure cloud storage framework effectively combines chaos-based cryptography, authenticated encryption, and blockchain-based verification, thereby delivering a comprehensive security solution for cloud environments. The combination of Lorenz 3D chaotic key generation, AES-256-GCM encryption, and PBFT-based blockchain verification achieves multiple security objectives, including confidentiality, integrity, authentication, auditability, and fault-tolerance.

The randomness analysis of the generated chaotic keys validates the effectiveness of the Lorenz chaotic system in cryptographic key generation. The Shannon entropy value was 7.99, the bit distribution was balanced, the autocorrelation values were close to zero, and the NIST SP 800-22 test results were positive, which shows that the generated keys have good randomness properties and low statistical bias. These properties greatly improve the unpredictability and resistance to brute-force and statistical attacks.

The robustness of the encryption layer can also be validated using ciphertext analysis. The ciphertext had near-ideal entropy, low chi-square deviation, negligible auto-correlation, and a mean avalanche effect of approximately 50%, indicating good diffusion and confusion properties. The findings show that the proposed encryption scheme successfully hides the pattern of the plaintext and reduces information leakage, which enhances its differential and statistical cryptanalysis resistance.

The performance evaluation demonstrates that the latencies for encryption, upload, and total processing grow steadily as the file size increases, as expected for cryptographic operations on larger datasets. Although the integration of chaotic key generation and blockchain verification adds some computational overhead, the measured PBFT consensus latency remains very low (approximately 0.01 sec), suggesting that integrity verification can be performed efficiently without a significant impact on the overall system performance. As the file sizes increase, the throughput decreases; however, the framework remains efficient while offering significantly improved security promises.

The comparative analysis shows that the conventional cloud storage has a higher throughput and lower latency but lacks some advanced security capabilities, such as decentralized integrity verification, blockchain auditability, Byzantine fault tolerance, and chaos-enhanced cryptographic randomness. The proposed framework, on the other hand, has a

relatively small overhead and provides much greater protection. The blockchain layer provides immutability to audit records and tamper-proof features, whereas the PBFT consensus mechanism guarantees dependable operation even when dealing with malicious validator nodes.

The security evaluation and results under various attack scenarios demonstrate the effectiveness of the proposed design. The combination of the large key space of (2^{256}), good avalanche behavior, authenticated encryption, dual-hash verification and immutability of the blockchain offers significant protection against unauthorized access, brute force attacks, ciphertext manipulation, replay attacks, known-plaintext attacks, chosen-ciphertext attacks, and Byzantine failures. Based on these findings, a defense-in-depth security architecture is achieved in the proposed framework, which is applicable to sensitive cloud-based applications.

However, the present implementation was tested in a local consortium blockchain setup with a small number of validator nodes. Hence, the reported PBFT latency and throughput numbers might vary in geographically distributed cloud deployments, where network delays and communication overhead have a greater impact. Additionally, the communication complexity of the PBFT consensus is $(O(n^2))$, potentially impacting scalability as the number of validators increases.

Overall, the results demonstrate that the proposed framework effectively achieves a trade-off between security and performance through the use of chaos-based key generation, authenticated encryption, and blockchain verification of integrity. The architecture is well-suited for high-security needs and verifies the integrity of data, such as healthcare information systems, enterprise cloud storage platforms, financial services, and Internet of Things (IoT) applications.

A. Effect of Euler Integration on Chaotic Key Quality

The Lorenz system was numerically integrated using the Euler method; however, the statistical evaluation results showed that the chaotic sequences generated by this system were highly random throughout the experiments. The entropy value of 7.99 bits, successful NIST SP 800-22 test results, and low autocorrelation indicate that the generated key sequences were statistically uniform in the range of iterations tested and file sizes. Based on these results, it seems that the numerical errors caused by the Euler method did not significantly affect the cryptographic quality of the generated keys with the experimental settings. However, future studies could explore the use of higher-order integration methods, such as the fourth-order Runge–Kutta algorithm, to further enhance the long-term numerical accuracy and chaotic trajectory stability.

B. Limitations of the PBFT Performance Evaluation

The PBFT performance evaluation was performed on a four-node consortium blockchain prototype. Hence, the measured latency and throughput data must be considered as prototype-scale numbers under ideal experimental conditions. Although the results show the feasibility of integrating PBFT with the proposed cloud security framework, they do not completely reflect the communication overhead that can occur in a larger consortium blockchain application. As the number of validator nodes grows, the network traffic and consensus latency of PBFT are expected to grow as well, because its message complexity is $(O(n^2))$ of the number of nodes. Therefore, these reported performance metrics are not necessarily applicable to large-scale cloud environments. The framework will be tested for scalability in future work across geographically distributed nodes and larger networks of validators to validate it in realistic deployment scenarios.

XII. Conclusion

This study proposed a secure integrated cloud storage system that integrates the Lorenz 3D chaotic key generation method, AES-256-GCM authenticated encryption, and blockchain verification technology based on PBFT. The proposed system enhances data confidentiality, integrity, auditability, and Byzantine fault tolerance by storing encrypted files in the cloud and integrity metadata on the blockchain. The experimental outcomes demonstrated excellent cryptographic performance, such as a nearly perfect entropy of 7.99, an avalanche effect of 50.05%, low autocorrelation, and successful randomness validation. The framework also proved to be resistant to brute-force, statistical, differential, replay, tampering, known-plaintext, chosen-ciphertext, and Byzantine attacks. Although the evaluation was performed in a local prototype setting, the outcome of the experiment validates the proposed framework of providing a secure and efficient multilayer solution for cloud storage applications.

XIII. Future Work

This work can be extended in the future by the addition of post-quantum cryptographic techniques and deployed in a real-world, large-scale distributed cloud system with enhancements for performance and scalability.

AI Usage Disclaimer

The authors acknowledge the use of artificial intelligence (AI)-assisted tools for language refinement, grammar correction, formatting assistance, and improving the clarity of technical explanations during manuscript preparation. AI tools were not used to generate research results, experimental data, scientific conclusions, or core technical contributions. All system designs, implementations, analyses, and validations were independently conducted and verified by the authors.

TABLE IX. List of Notations and Abbreviations

Symbol / Notation	Description
P	Plaintext file
C	Ciphertext
K_{AES}	256-bit AES encryption key
IV	Initialization Vector (Nonce)
AAD	Associated Authenticated Data
T	Authentication tag generated by GCM
H_{plain}	SHA-256 hash of plaintext
H_{cipher}	SHA-256 hash of ciphertext
J_0	Initial counter block in GCM

J_i	Counter block for encryption round i
S_i	AES-generated keystream block
P_i	Plaintext block i
C_i	Ciphertext block i
H	GHASH hash subkey
G	GHASH output value
x, y, z	Lorenz chaotic state variables
x_0, y_0, z_0	Initial conditions of Lorenz system
σ	Lorenz system parameter (Prandtl number)
ρ	Lorenz system parameter (Rayleigh number)
β	Lorenz system parameter
h	Numerical integration step size
N	Number of chaotic iterations
B	Generated chaotic bitstream
K	Generated chaotic secret key
SHA-256	Secure Hash Algorithm (256-bit)
AES-256-GCM	Advanced Encryption Standard with Galois/Counter Mode
PBFT	Practical Byzantine Fault Tolerance
Tx	Blockchain transaction
FileID	Unique file identifier
Timestamp	File upload time recorded on blockchain
Smart Contract	Blockchain-based access control logic
Validator Node	Blockchain participant responsible for consensus
PRE-PREPARE	First PBFT consensus phase
PREPARE	Second PBFT consensus phase
COMMIT	Final PBFT consensus phase

REFERENCES

[1] S. S. Gill and R. Buyya, "A taxonomy and future directions for sustainable cloud computing: 360 degree view," ACM Comput. Surv., vol. 51, no. 5, 2018, doi: 10.1145/3241038.

[2] R. Buyya et al., "A manifesto for future generation cloud computing: Research directions for the next decade," ACM Comput. Surv., vol. 51, no. 5, Sep. 2019, doi: 10.1145/3241737.

[3] P. Yang, N. Xiong, and J. Ren, "Data Security and Privacy Protection for Cloud Storage: A Survey," IEEE Access, vol. 8, pp. 131723–131740, 2020, doi: 10.1109/ACCESS.2020.3009876.

[4] NIST, "Advanced Encryption Standard (AES)," May 2023. doi: 10.6028/NIST.FIPS.197-upd1.

[5] D. P. Leech, S. Ferris, and J. T. Scott, "The economic impacts of the advanced encryption standard, 1996?2017," Gaithersburg, MD, Aug. 2018. doi: 10.6028/NIST.GCR.18-017.

[6] S. Rehman, N. Talat Bajwa, M. A. Shah, A. O. Aseeri, and A. Anjum, "Hybrid aes-ecc model for the security of data over cloud storage," Electronics (Switzerland), vol. 10, no. 21, Nov. 2021, doi: 10.3390/electronics10212673.

[7] N. S. Musa, T. Murugan, N. A. N, and N. M. Mirza, "Research study on securing the cloud: utilizing conventional and blockchain-based access control mechanisms," Dec. 01, 2025, Springer Science and Business Media Deutschland GmbH. doi: 10.1186/s13677-025-00803-3.

[8] W. Lu, C. Jin, J. Wang, X. Liu, J. Liu, and Z. Zhai, "A novel image encryption scheme using 3D chaotic maps with Josephus permutation and dynamic diffusion," Journal of King Saud University - Computer and Information Sciences, vol. 37, no. 8, Oct. 2025, doi: 10.1007/s44443-025-00284-z.

[9] Z. Rahman, X. Yi, M. Billah, M. Sumi, and A. Anwar, "Enhancing AES Using Chaos and Logistic Map-Based Key Generation Technique for Securing IoT-Based Smart Home," Electronics (Switzerland), vol. 11, no. 7, Apr. 2022, doi: 10.3390/electronics11071083.

[10] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008, doi: 10.1007/s10838-008-9062-0.

[11] P. Sharma, R. Jindal, and M. D. Borah, "Blockchain-based decentralized architecture for cloud storage system," Journal of Information Security and Applications, vol. 62, Nov. 2021, doi: 10.1016/j.jisa.2021.102970.

[12] P. J. Sun, "Security and privacy protection in cloud computing: Discussions and challenges," Journal of Network and Computer Applications, vol. 160, no. August 2019, p. 102642, 2020, doi: 10.1016/j.jnca.2020.102642.

[13] M. N. Birje, P. S. Challagidad, R. H. Goudar, and M. T. Tapale, "Cloud computing review: concepts, technology, challenges and security," 2017.

[14] M. B. Qureshi et al., "Encryption Techniques for Smart Systems Data Security Offloaded to the Cloud," Symmetry (Basel)., vol. 14, no. 4, Apr. 2022, doi: 10.3390/sym14040695.

[15] R. Ganesh, B. U. I. Khan, A. R. Khan, and A. Bin Kamsin, "A panoramic survey of the advanced encryption standard: from architecture to security analysis, key management, real-world applications, and post-quantum challenges," Int. J. Inf. Secur., vol. 24, no. 5, Oct. 2025, doi: 10.1007/s10207-025-01116-x.

[16] P. Sharma, R. Jindal, and M. D. Borah, "Blockchain Technology for Cloud Storage: A Systematic Literature Review," ACM Comput. Surv., vol. 53, no. 4, 2020, doi: 10.1145/3403954.

[17] C.-H. Lin, G.-H. Hu, C.-Y. Chan, J.-J. Yan, and J. Chaos-Based, "Chaos-Based Synchronized Dynamic Keys and Their Application to Image Encryption with an Improved AES Algorithm," 2021, doi: 10.3390/app.

[18] B. Zhang and L. Liu, "Chaos-Based Image Encryption: Review, Application, and Challenges," Jun. 01, 2023, MDPI. doi: 10.3390/math11112585.

[19] M. A. Khan, "New Image Encryption Scheme Using Chaotic Maps," in ACM International Conference Proceeding Series, Association for Computing Machinery, Aug. 2019. doi: 10.1145/3387168.3389111.

[20] A. Arab, M. J. Rostami, and B. Ghavami, "An image encryption method based on chaos system and AES algorithm," Journal of Supercomputing, vol. 75, no. 10, pp. 6663–6682, Oct. 2019, doi: 10.1007/s11227-019-02878-7.

[21] A. Alghuried, M. Alkinoon, M. Mohaisen, A. Wang, C. Zou, and D. Mohaisen, "Blockchain Security and Privacy: Threats, Challenges, Applications, and Tools," Distributed Ledger Technologies, vol. 5, no. 1, Dec. 2025, doi: 10.1145/3716323.

[22] Z. Wenhua, F. Qamar, T. A. N. Abdali, R. Hassan, S. T. A. Jafri, and Q. N. Nguyen, "Blockchain Technology: Security Issues, Healthcare Applications, Challenges and Future Trends," Feb. 01, 2023, MDPI. doi: 10.3390/electronics12030546.

[23] X. Li, P. Jiang, T. Chen, X. Luo, and Q. Wen, "A Survey on the Security of Blockchain Systems," 2020.

[24] J. Singh et al., "A Systematic Review of Blockchain, AI, and Cloud Integration for Secure Digital Ecosystems," Dec. 01, 2025, Springer Science and Business Media B.V. doi: 10.1007/s44227-025-00072-1.

[25] C. Yu, N. Mei, C. Du, and H. Luo, "Blockchain Data Scalability and Retrieval Scheme Based on On-Chain Storage Medium for Internet of Things Data," Electronics (Switzerland), vol. 12, no. 6, Mar. 2023, doi: 10.3390/electronics12061454.

[26] Y. Luo, M. Liu, J. Wang, C. Yuan, and T. Liu, "When Secure Data Sharing Meets Blockchain: Overview, Challenges and Future Prospects," in ACM International Conference Proceeding Series, Association for Computing Machinery, Mar. 2022, pp. 1–8. doi: 10.1145/3532640.3532641.

[27] M. N. Alatawi, "Blockchain-Driven Smart Contracts for Advanced Authorization and Authentication in Cloud Security," Electronics (Switzerland), vol. 14, no. 15, Aug. 2025, doi: 10.3390/electronics14153104.

[28] Z. Yao, Y. Fang, H. Pan, X. Wang, and X. Si, "A secure and highly efficient blockchain PBFT consensus algorithm for microgrid power trading," Sci. Rep., vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-58505-w.

[29] J. Liu, X. Deng, W. Li, and K. Li, "CG-PBFT: an efficient PBFT algorithm based on credit grouping," Journal of Cloud Computing, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13677-024-00643-7.

[30] F. Nujumudeen et al., "A hybrid chaos-based cryptographic framework for lightweight IoT security: enhancing efficiency and security in low-power devices," Peer. Peer. Netw. Appl., vol. 19, no. 2, p. 53, Feb. 2026, doi: 10.1007/s12083-025-02188-1.

[31] M. Dworkin, "Recommendation for Block Cipher Modes of Operation: Galois/Counter Mode (GCM) and GMAC," NIST Special Publication 800-38D, Nov. 2007, doi: 10.6028/NIST.SP.800-38d.

[32] A. Amanat, M. Rizwan, C. Maple, Y. Bin Zikria, A. S. Almadhor, and S. W. Kim, "Blockchain and cloud computing-based secure electronic healthcare records storage and sharing," Front. Public Health, 2022, [Online]. Available: https://www.frontiersin.org/

[33] T. Zhang, C. Wang, and M. I. U. Chandrasena, "Blockchain-assisted data sharing supports deduplication for cloud storage," Conn. Sci., vol. 35, no. 1, 2023, doi: 10.1080/09540091.2023.2174081.

[34] R. Golla Bala and S. Gnanavel, "BC2P-1305: an enhanced data security in cloud computing network using blockchain based ChaCha20-Poly1305 cryptography," Frontiers in Blockchain, vol. 8, 2025, doi: 10.3389/fbloc.2025.1636056.

Comparative Evaluation of Traditional and Transformer-Based Models for Risk-Level Classification of Uzbek Telegram Messages

Feruzakhon A. Qoyliyeva[1], Ozod J.Babomuradov[2], Akmal A. Savurbayev[3]
Tashkent State Agrarian University, Tashkent, Uzbekistan[1]
Head of Department for Coordination of Digitalization and Artificial Intelligence Development,
Administration of the President of the Republic of Uzbekistan, Jizzakh, Uzbekistan[2]
Department of Financial Technologies, Digitalization and Artificial Intelligence,
Presidential Administration of the Republic of Uzbekistan, Tashkent, Uzbekistan[3]

***Abstract*—The rapid growth of Telegram-based communication has increased the dissemination of harmful and risky content, particularly in low-resource languages such as Uzbek. This study investigates the automatic classification of Uzbek Telegram messages according to risk level using both traditional machine learning and transformer-based models. A dataset consisting of 10,000 real Telegram messages was collected and manually annotated into two classes: Safe and Dangerous. To improve data quality and consistency, preprocessing techniques including URL removal, emoji normalization, stop-word filtering, and script unification were applied. The study compares the performance of TF-IDF + Logistic Regression, FastText, mBERT, and XLM-RoBERTa for harmful content detection in Uzbek Telegram texts. Experimental results show that transformer-based models significantly outperform traditional approaches. Among all evaluated models, XLM-RoBERTa achieved the highest performance, with an Accuracy of 91.2%, a precision of 90.8%, a recall of 91.5%, and an F1-score of 91.1%, while mBERT achieved an Accuracy of 84.9% and an F1-score of 84.6%. The results demonstrate the effectiveness of contextual transformer architectures for identifying harmful content in low-resource language environments. The findings confirm that transformer-based models provide a reliable solution for automatic risk-level classification of Uzbek social media texts and can support practical applications in content moderation, information security, and social media monitoring systems.**

***Keywords*—Uzbek language; telegram messages; risk-level classification; harmful content detection; machine learning; transformer models; mBERT; XLM-RoBERTa**

I. Introduction

In recent years, the rapid growth of textual data generated through social media platforms has led to emerging challenges related to information security, public order, and digital culture. In particular, short, informal, and often unmoderated content shared on platforms such as Telegram and YouTube has made the timely detection of harmful content a critical issue [1–2].

Traditional rule-based approaches and classical machine learning methods such as Naive Bayes, Support Vector Machine, and Logistic Regression often fail to capture the deep semantic meaning of such complex textual data [3]. This limitation becomes even more pronounced for the Uzbek language due to the coexistence of Latin and Cyrillic scripts, widespread use of slang, abbreviations, and emojis [4].

The primary objective of this study is to perform a comparative analysis of traditional machine learning and transformer-based models for classifying Uzbek social media texts according to risk levels and to identify the most effective approach.

The exponential growth of textual content on social media has intensified the need for automated detection of hate speech, toxic expressions, user sentiment, and content safety. Recent studies indicate that deep learning techniques, particularly transformer-based architectures, significantly outperform traditional approaches in these tasks [5–6].

Several studies have demonstrated that multimodal approaches (text + image) can further improve hate speech detection performance. For instance, models based on RoBERTa and Swin Transformer have achieved high F1 scores in meme classification tasks, showing that combining textual and visual features yields better performance than single-modality approaches [7–8].

Transformer models such as BERT, RoBERTa, mBERT, and XLM-R have consistently demonstrated superior performance compared to classical machine learning methods (e.g., SVM, Logistic Regression, Random Forest), even in multilingual and imbalanced dataset scenarios. In particular, RoBERTa and XLM-R show stable results in low-resource language settings [9].

Research conducted for the Uzbek language indicates that monolingual pretrained models such as BERTbek can outperform multilingual models in certain NLP tasks. However, challenges such as script variation (Latin/Cyrillic) and limited corpus size may negatively affect model accuracy [10].

In the context of social media and user-generated content analysis, sentiment analysis and aspect-based classification tasks have also been widely studied. While classical machine learning models achieve moderate performance, transformer-based approaches have demonstrated significantly higher accuracy in sentiment detection and categorization tasks [11–

14]. Furthermore, incorporating non-standard elements such as emojis has been shown to improve model performance [15].

Overall, existing studies suggest that transformer-based models represent the most effective approach for analyzing social media text. However, challenges such as the scarcity of annotated datasets in Uzbek, the complexity of manual labeling, and the difficulty of capturing context-dependent expressions remain unresolved [12,16]. Therefore, this study aims to explore the effectiveness of modern transformer models for the automatic analysis of Uzbek social media texts.

Although previous studies have shown competitive results for BERTbek, this study focuses on comparing traditional machine learning approaches with multilingual transformer models. Evaluation of monolingual Uzbek transformer models is left for future work.

II. Theoretical Background and Proposed Approach

In this study, a real-world dataset consisting of Uzbek social media texts was collected from the Telegram platform. Telegram is characterized by rapid information dissemination, high user engagement, and diverse content types, making it a suitable source for studying both harmful and safe content.

The dataset was compiled from public Telegram channels, discussion groups, and comments under posts covering a variety of topics, including social, everyday, political, religious, and informational content. In total, 10,000 textual messages were collected and analyzed from an information security perspective.

Within the scope of this research, the texts were categorized into two classes based on their semantic content:

- 0 – Safe: Neutral or positive messages that do not pose a threat to public safety.
- 1 – Dangerous: Messages containing hate speech, violence, provocation, or content that may threaten societal safety.

To prepare the dataset for model training, a multi-stage preprocessing pipeline was applied. In the first stage, irrelevant elements such as URLs, HTML tags, excessive punctuation, and special characters were removed. This step helped eliminate noise and reduce non-informative features.

In the next stage, emojis and emoticons were normalized into standard tokens or corresponding textual representations to preserve emotional context within the text. Additionally, due to the mixed usage of Latin and Cyrillic scripts in Uzbek, all texts were converted into a unified format.

Subsequently, stop words (low-information auxiliary words) were removed to enhance the importance of semantically meaningful keywords. After that, tokenization was performed, and the processed texts were transformed into input representations compatible with transformer-based models.

The labeling process was conducted manually through expert annotation. Each message was carefully analyzed based on its semantic meaning and assigned an appropriate risk label. This approach ensured higher reliability compared to automatic labeling methods and improved the quality of model training.

As a result, a cleaned and annotated dataset was prepared for training and evaluating transformer-based models.

III. Method

In this study, various machine learning and deep learning models with different architectures were comparatively analyzed to evaluate their effectiveness in classifying Uzbek social media texts by risk level. The selected models range from traditional approaches to modern transformer-based architectures.

A. Traditional and Neural-Based Models

In the initial stage, a TF-IDF + Logistic Regression model was applied. Texts were transformed into vector representations using TF-IDF, followed by classification using Logistic Regression. This approach served as a baseline model for comparison with transformer-based methods.

Next, the FastText model was employed. FastText operates at the subword level, making it particularly effective for morphologically rich languages. This allowed the model to capture suffixes and various word forms in the Uzbek language.

Due to computational and time constraints, each model was evaluated using a single train-test split. Future work will include multiple runs with different random seeds to further assess model stability and reproducibility.

B. Transformer-Based Models

The main focus of this study was on transformer-based architectures. In particular, the mBERT model was utilized due to its training on large multilingual corpora, enabling contextual understanding of Uzbek text. mBERT evaluates each word within its surrounding context, achieving higher accuracy compared to traditional embedding methods.

Additionally, the XLM-RoBERTa (XLM-R) model was applied. This model is trained on massive multilingual datasets and is known for its strong performance in low-resource languages. XLM-R demonstrated superior ability in capturing the semantic meaning of Uzbek texts, including slang, abbreviations, and informal expressions, outperforming other models in this study.

C. Training Procedure and Hyperparameters

All transformer-based models were fine-tuned using pre-trained weights and adapted to the specific dataset used in this study. The fine-tuning process was conducted using the following hyperparameters:

- Learning rate: 2e-5
- Batch size: 16
- Number of epochs: 4
- Loss function: Cross-Entropy

These hyperparameters were selected based on experimental validation to ensure model stability and generalization capability.

D. Dataset Split and Evaluation Strategy

The dataset was randomly divided into 80% training and 20% testing subsets. This approach enabled an unbiased evaluation of model performance on unseen data.

All models were trained under identical conditions, and their performance was evaluated using standard classification metrics, including:

- Accuracy
- Precision
- Recall
- F1-score

This ensured a fair and consistent comparison of model effectiveness[17-19].

As a result, the comparative analysis clearly demonstrated the superiority of transformer-based models over traditional approaches and provided a solid methodological foundation for selecting the most effective model.

IV. Results and Discussion

This section presents the experimental results of the models applied for classifying Uzbek social media texts according to risk levels. All models were trained under the same conditions and evaluated on the test dataset. The results provide an objective comparison of overall performance and class-wise effectiveness.

A. Overall Model Performance

Table I summarizes the evaluation results of the applied models based on the key performance metrics: Accuracy, Precision, Recall, and F1-score.

TABLE I. Performance Comparison of Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
TF-IDF + Logistic Regression	71.4	70.2	69.8	69.9
FastText	78.6	77.9	78.1	78.0
mBERT	84.9	84.2	85.1	84.6
XLM-R	91.2	90.8	91.5	91.1

All Precision, Recall, and F1-score values are reported as weighted averages calculated on the test dataset.

B. Results Analysis and Discussion

As shown in Table I, the traditional TF-IDF + Logistic Regression model provides baseline performance; however, its ability to capture deep semantic relationships remains limited. Although the FastText model achieves improved results due to its subword-level representation, its overall performance is still inferior compared to transformer-based models.

The mBERT model demonstrates significantly better performance by leveraging multilingual contextual representations. Nevertheless, the highest performance across all evaluation metrics is achieved by the XLM-RoBERTa model, which consistently outperforms all other models.

C. Threshold Analysis

The effect of classification threshold selection was evaluated using the XLM-RoBERTa model. A threshold value of 0.7 was chosen to increase prediction reliability and reduce false-positive classifications. Under this setting, the model classifies a message as dangerous only when a higher confidence level is achieved[20-21].

Experiments conducted on the dataset of 10,000 messages demonstrated that the threshold of 0.7 improves classification stability and reliability. Although fewer messages are classified as dangerous, the precision of harmful content detection increases, making the system more suitable for practical applications such as content moderation and information security.

For threshold validation, a balanced subset of 200 messages (100 safe and 100 dangerous) was selected from the main dataset. This subset was used to evaluate the effect of threshold adjustment on model behavior and prediction confidence.

The results indicate that a higher threshold reduces false positives while maintaining strong classification performance. In particular, XLM-RoBERTa demonstrated stable performance in distinguishing between Safe and Dangerous messages. The high Recall observed for the Dangerous class suggests that the model effectively minimizes the risk of missing harmful content, which is essential for real-world monitoring and security systems.

Overall, the combination of a 10,000-message dataset and a threshold of 0.7 provides a reliable balance between precision and recall, enabling accurate identification of harmful content in Uzbek social media texts.

D. Error Analysis

The analysis of misclassified instances revealed that most errors occurred in texts containing sarcasm, irony, or implicit meanings. Additionally, very short or context-limited messages posed challenges for all models in determining the correct risk level.

Despite these challenges, the XLM-R model maintained relatively stable performance even in complex cases. This can be attributed to its training on large-scale multilingual corpora and its ability to capture deep contextual relationships.

E. Summary of Results

Overall, the experimental results clearly demonstrate that transformer-based models—particularly XLM-RoBERTa—achieve high effectiveness in classifying Uzbek social media texts by risk level. The model exhibits high accuracy, robustness, and generalization capability, making it a suitable solution for practical deployment [21-22].

F. Discussion

The findings of this study highlight the superior performance of transformer-based models in classifying Uzbek social media texts by risk level. The experimental results confirm that deep learning approaches significantly outperform traditional machine learning methods.

The baseline TF-IDF + Logistic Regression approach relies on surface-level textual features and fails to adequately capture contextual meaning. This limitation becomes particularly evident in emotionally rich or informal texts. The FastText model partially addresses this issue through subword representations; however, it still lacks the ability to fully capture long-range semantic dependencies.

The mBERT model demonstrates improved performance due to its multilingual contextual understanding. However, architectural limitations and the presence of cross-lingual noise slightly restrict its effectiveness.

The best performance is achieved by the XLM-RoBERTa model. This can be explained by its training on large-scale multilingual datasets and its ability to capture deep semantic relationships, even in low-resource languages. Furthermore, XLM-R effectively handles Uzbek-specific challenges such as slang, abbreviations, emojis, and mixed Latin-Cyrillic scripts.

Class-wise analysis indicates that high Recall in the Dangerous category is particularly important in practice, as it minimizes the risk of missing harmful content. Misclassifications between the Safe and Dangerous classes are mainly caused by ambiguous, context-dependent, or borderline messages that are difficult to interpret correctly. Despite these challenges, the transformer-based models demonstrated strong performance in distinguishing between the two classes.

At the same time, the study reveals several limitations. Even advanced transformer models struggle with sarcasm, irony, and implicit meanings in short texts. This suggests the need for future research focusing on deeper contextual modeling and multimodal approaches [23-25].

Overall, the results confirm that dataset expansion, improved preprocessing, and appropriate model selection play a crucial role in achieving high-performance text classification systems. The study successfully achieves its objective of developing an accurate and robust risk-level classification approach for Uzbek social media texts.

V. Conclusion

This study presents a comprehensive analysis of risk-level classification of Uzbek social media texts using various machine learning and deep learning models. The experimental results clearly demonstrate that transformer-based models—especially XLM-RoBERTa—significantly outperform traditional approaches.

The findings highlight that dataset expansion, optimized preprocessing, and proper model selection have a substantial impact on performance. The XLM-R model achieved high Accuracy, Precision, Recall, and F1-score, enabling more accurate semantic classification of informal, short, and emotionally rich Uzbek texts. In particular, the high Recall in detecting harmful content makes this model highly suitable for practical applications in information security systems.

The results indicate that even for low-resource languages such as Uzbek, it is possible to develop high-quality automated text analysis systems by applying appropriate methodologies and modern transformer architectures. This has important implications for social media monitoring, early warning systems, and ensuring information security in digital environments.

However, the study also identifies certain limitations. Detecting sarcasm, irony, and implicit meanings remains a challenging task. Additionally, very short and context-limited texts continue to pose difficulties for classification models.

Future research directions include:

- Developing real-time harmful content monitoring systems.
- Applying multimodal analysis (text + image + video) for risk detection.
- Expanding the dataset and improving class balance.
- Incorporating contextual and dialogue-based models for detecting sarcasm and implicit meanings.

The results of this study provide a strong foundation for further research in intelligent analysis of Uzbek social media content and open opportunities for real-world implementation.

References

[1] E. Kuriyozov et al., “BERTbek: A pretrained language model for Uzbek,” in Proc. SIGUL @ LREC-COLING, 2024.

[2] B. Mansurov and A. Mansurov, “UzBERT: Pretraining a BERT model for Uzbek,” arXiv preprint arXiv:2108.09814, 2021.

[3] E. Kuriyozov et al., “Text classification dataset and analysis for Uzbek language,” arXiv preprint arXiv:2302.14494, 2023.

[4] S. G. Matlatipov et al., “Aspect-based sentiment analysis for the Uzbek language (UzABSA dataset),” in Proc. SIGUL @ LREC-COLING, 2024.

[5] S. Allanazarova and D. Elova, “Uzbek texts sentiment analysis: Database development,” in Proc. ScitePress, 2023.

[6] K. Madatov et al., “TF-IDF-based classification of Uzbek educational texts,” Applied Sciences, vol. 15, no. 19, Art. no. 10808, 2025.

[7] X. Wang, “CLTL @ Multimodal hate speech event detection,” in Proc. CASE @ EACL, 2024.

[8] S. Thapa, “Extended multimodal hate speech event detection during Russia–Ukraine war,” in Proc. CASE, 2024.

[9] T. Mandl et al., “HASOC 2024: Hate speech identification in English and Bengali,” in Proc. CEUR Workshop, 2024.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451.

[11] S. Matlatipov et al., “UzABSA: Aspect-based sentiment analysis for Uzbek,” in Proc. SIGUL, 2024.

[12] A. Taniberdiyev, “Ijtimoiy tarmoq matnli yozishmalarini tasniflash texnologiyalari,” CyberLeninka, 2024.

[13] Y. Yusufu et al., “A case study of the Uzbek language Google Play reviews,” Information Processing & Management, 2025.

[14] HuggingFace, “Uzum Market sentiment dataset,” 2024–2025.

[15] I. Rabbimov, I. Mporas, V. Simaki, and S. Kobilov, “Investigating the effect of emoji in opinion classification of Uzbek movie review comments,” 2020.

[16] S. Matlatipov, J. Rajabov, E. Kuriyozov, and G. Matlatipov, “Uzbek sentiment analysis based on local restaurant reviews,” 2022.

[17] O. J. Babomuradov and F. A. Qo‘yliyeva, “Analysis and comparison of methods for evaluating social media messages,” in Proc. Int. Sci. Pract. Conf. Scientific-Technical and Sociocultural Problems of Modern Society, Jizzakh, Uzbekistan, 2025, pp. 114–118.

[18] J. Sosa et al., “Multimodal pipeline for collection of misinformation data from Telegram,” in Proc. LREC, 2022.

[19] P. Kapil et al., “A survey on combating hate speech through detection and countermeasures,” in Proc. ICON (ACL), 2024.

[20] S. Al-Saqqa et al., “A survey of hate speech detection for Arabic social media,” Procedia Computer Science, vol. 238, 2024.

[21] J. Opitz et al., “A closer look at classification evaluation metrics and a guide for practitioners,” Trans. ACL, 2024.

[22] A. Abdullayev et al., “Dialect-sensitive sentiment analysis for Uzbek news,” in Proc. Conf., 2025.

[23] J. Rajabov, “Analysis and retraining of the BERT model for the Uzbek language: Methods and results,” in Proc. Int. Conf. Educational Discoveries and Humanities, 2024.

[24] UnitaryAI, “Detoxify: Trained models and code to predict toxic comments,” GitHub repository, 2020–2025.

[25] A. Banerjee et al., “A comprehensive survey on transformer-based machine translation,” ACM Computing Surveys, 2025.

MedChain: A Privacy-Preserving Framework for Scalable Electronic Health Record Sharing on Blockchain and InterPlanetary File System

S. Venkateswaran[1], N. Vijayaraj[2]
Research Scholar[1], Professor[2]
Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India[1, 2]

***Abstract*—The secure and scalable sharing of electronic health records (EHRs) remains a fundamental challenge in modern health-care systems due to conflicting requirements of privacy, regulatory compliance (HIPAA, GDPR), and real-time clinical access. Existing blockchain-based solutions suffer from three critical limitations, including static access control policies that cannot adapt to emergency scenarios, complete transparency of policy evaluation leading to attribute leakage, and linear degradation of throughput with policy complexity. The proposed MedChain, a framework introducing four novel contributions: Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE) enabling time-bound and emergency breakout access without re-encrypting underlying data; Zero-Knowledge Policy Verification (ZK-PV) allowing smart contracts to validate access requests without revealing sensitive attribute values; Adaptive EHR Sharding that dynamically adjusts IPFS shard sizes based on access frequency, reducing cold-start latency by 64%; and Layered Revocation Cascade providing granular revocation at patient, provider, and record-type levels within 2.3 seconds. Using the MIMIC-III benchmark (46,520 patients, 100,000 access traces), we demonstrate: 99.4% storage reduction vs. pure blockchain, 0.95s median latency (43% improve-ment over prior work), 2,450 requests/second throughput with 150 IPFS nodes, and provable security under the q-DBDH assumption. Comprehensive comparisons with MedRec, FHIRChain, Ancile, and PriHac confirm MedChain's superiority across all metrics.**

***Keywords*—Blockchain; electronic health records; IPFS; dynamic attribute-based proxy re-encryption; zero-knowledge proofs; adaptive sharding; MIMIC-III**

I. Introduction

The fast-paced digitalization of the healthcare industry has led to an increased use of electronic health records (EHR) to record and share patient medical data. The use of EHRs increases the accuracy of diagnoses and reduces time in treatments while facilitating communication between hospitals, labs, insurance companies, and emergency departments [1]. Nevertheless, the secure exchange of EHR data is one of the biggest challenges due to privacy issues, compliance regulations, and the need for immediate access during emergencies. Healthcare information is very sensitive in nature and is supposed to be compliant with international standards, such as HIPAA and GDPR [2].

Traditional centralized EHR solutions like Epic Systems and Oracle Health depend entirely on third-party servers to store and process their data, making them prone to attacks [3][4]. The large-scale breaches that have occurred in recent times have shown that centralized EHR solutions are inadequate for protecting health records from any breach and compromise.

The use of blockchain technology is considered a viable approach owing to its decentralized structure, non-repudiation, openness, and traceability [5]. The removal of reliance on centralized entities facilitates trust building among healthcare entities while ensuring tamper-proof access management [6]. Blockchain-driven systems in the field of healthcare, such as MedRec, FHIRChain, Ancile, and PriHac, have sought to address the issue of secure EHR sharing [7]. While these systems offer some benefits, there are numerous obstacles that restrict their implementation. Blockchain offers decentralization and immutability, but prior systems fail on three fronts:

- Static Policies: The majority of current systems use static access control policies whereby permission is configured before encryption [8][9]. Any alteration to the policy, for instance, in case of temporary access by specialists, emergencies involving "break glass" access, or a change in the provider's role, necessitates policy reconfiguration as well as encryption of the patient's data [10].
- Attribute Leakage: Access control policies in traditional blockchain frameworks are assessed using smart contracts within the network. Due to the public nature of the blockchain, private data like patient condition, doctor specialty, department of the hospital, and emergencies may get disclosed, leading to potential privacy violations [11][12].
- Performance and Scalability Degradation: The performance and scalability of blockchain-based access authentication and proxy re-encryption suffer immensely with increasing numbers of users, attributes, and access queries. Current systems demonstrate either linear or quadratic growth in computation, leading to poor response time and low processing speeds [13] [14]. It is impractical to rely on such mechanisms in large-scale hospitals that require continuous access in real time. Such challenges emphasize the immediate necessity for

an EHR sharing architecture that ensures security, scalability, and privacy [15][16][17].

A. Novel Contributions Overview

In this regard, this study introduces the concept of MedChain, a framework for secure EHR sharing using blockchain that employs dynamic attribute-based proxy re-encryption, zero-knowledge policy verification, adaptive EHR sharding, and layered revocation cascade.

The architecture incorporates blockchain technology along with proxy re-encryption, zero-knowledge proofs, and distributed storage using IPFS. Unlike conventional systems, where the EHR file is stored on-chain, this solution involves maintaining encrypted medical records via distributed storage and storing immutable metadata and access rights within the blockchain.

The main innovation of MedChain is its capability to provide dynamic access control, privacy-preserving policy verification, intelligent storage management, and hierarchical revocation under one umbrella. Contrary to previous approaches, MedChain allows time-limited access control, emergency access control, and modification of attributes without re-encryption of the original EHRs.

TABLE I. NOVELTY PROPOSED MEDCHAIN COMPARED TO PRIOR WORKS

Feature	MedRec	FHIR Chain	Ancile	MedChain (Ours)
Dynamic AB-PRE				✓
Zero-knowledge policy verification				✓
Adaptive EHR sharding				✓
Layered revocation cascade	Partial		Partial	✓
Formal security with q-DBDH			Partial	✓
Benchmark on real dataset (MIMIC-III)			✓	✓
Emergency breakout access				✓
Time-bound keys		✓		✓

MedChain is thus a comprehensive system that includes dynamic policy management, privacy-preserved access control, storage optimization, and fine-grained revocation, as well as formal cryptographic security, as evident from Table I. The comparison clearly demonstrates that the proposed framework offers more comprehensive functionality and stronger formal security guarantees than previous approaches.

B. Technical Contributions

1) Dynamic attribute-based proxy re-encryption (DAB-PRE): A new form of DAB-PRE that can facilitate four types of access modes, namely, static authorization, time-limited access, emergency break-glass access, and dynamic attributes change access without the need to re-encrypt the existing EHRs.

2) Zero-Knowledge Policy Verification (ZK-PV): Zero-Knowledge Policy Verification (ZK-PV), which uses Bulletproofs to achieve zero knowledge, is proposed as a solution to enable smart contracts to verify access requests made by patients and providers without disclosing any private data.

3) Adaptive EHR Sharding (AES): A machine-learning-based sharding technique where access frequency is predicted and shard sizes dynamically adjusted, resulting in a 64% reduction in cold-start retrieval latency.

4) Layered Revocation Cascade (LRC): A hierarchy-based revocation mechanism is suggested that works on four levels: patient level, provider level, record type level, and individual record level. The mechanism ensures prompt revocation with a maximum propagation time of just 2.3 seconds.

5) Comprehensive evaluation: The suggested model is tested on the MIMIC-III standard data set containing 46,520 patients and 100,000 access logs. The findings demonstrate that there is a 99.4% decrease in storage space when compared to existing blockchain technology, an access latency time of 0.95 seconds, a throughput of 2,450 queries per second utilizing a 150-node IPFS network, and security assurance based on the q-DBDH problem.

C. Study Organization

The rest of the study is structured as follows. Section II describes the literature review, followed by an overview of recent EHR sharing systems based on blockchain technology. In Section III, the proposed MedChain framework is described along with the details of the system architecture. Section IV is about the cryptography and access control model. Performance evaluation and comparison analysis with benchmark data sets are presented in Section V. Security analysis, and its formal proof is presented in Section VI.

II. RELATED WORKS

The need for sharing Electronic Health Records securely has emerged as an important field of study since healthcare information needs to be kept confidential and accessible simultaneously by physicians, healthcare organizations, laboratories, and emergency services. Blockchain technology has become quite popular in healthcare due to the benefits, such as decentralization, transparency, immutability, and enhanced security, that it offers. Several researchers have integrated blockchain technology with ABE, PRE, IPFS, and ZKP schemes to enhance secure EHR sharing. There are many limitations in most of the existing approaches, such as static access control, privacy breaches, inefficiency in revocation, and scalability issues.

MedRec is one of the earliest blockchains integrated with EHRs. Blockchain was used for access management and audit logs, while the patient information itself was stored off-chain. It offered transparency and control over data for patients. Yet, it required trusted third-party servers and did not provide emergency access and dynamic updates of policies [18]. Similarly, FHIRChain utilizes blockchain technology along with FHIR for healthcare interoperability. The blockchain system enabled hospitals to exchange data references related to the health records of patients instead of keeping the actual data on the blockchain. Though interoperability was improved in the

process, the blockchain technology incurred a large amount of gas fees and did not have flexible access control [19]. Ancile is based on blockchain and attribute-based encryption (ABE) for fine-grained access control. Though the system offered a privacy-preserving framework for sharing healthcare data, changes in access policies led to the re-encryption of the original data at the computational cost [20].

SecureMed adopted Proxy Re-Encryption using blockchain smart contracts for delegated EHR access, which led to enhanced security authorization amongst healthcare providers. Nonetheless, it made a presupposition of having static trust in proxy servers and did not allow temporary and emergency access [24]. ChainHealth integrated blockchain, ABE, and IPFS in an effort to enhance scalability and decrease storage costs. Even though this provided enhanced security in storing information, there was no zero-knowledge proof in accessing the data [25].

Zero-Knowledge Proofs (ZKP) were used by ZK-Health to implement privacy preservation verification in blockchain systems in healthcare applications. This minimized the information leakage while validating the transactions. However, it only concentrated on preserving the privacy of transactions but lacked support for attribute-based access control for Electronic Health Records (EHRs) [26]. MedShard proposed an innovative method using machine learning for implementing dynamic sharding in blockchain and IPFS systems for healthcare applications. This enhanced the storage management process and minimized the retrieval delay [27].

TABLE II. LITERATURE SURVEY OF BLOCKCHAIN-BASED SECURE EHR SHARING SYSTEMS

Ref. No	Author / System	Techniques Used	Key Contributions	Limitations
[18]	Azaria et al. / MedRec	Blockchain + Off-chain Storage	Patient-centric EHR management and auditability	Trusted server dependency, no emergency access
[19]	Mettler et al. / FHIRChain	Blockchain + FHIR + Smart Contracts	Improved interoperability and secure healthcare data exchange	High gas cost, no dynamic access control
[20]	Jiang et al. / Ancile	Blockchain + ABE + Smart Contracts	Fine-grained access control and privacy preservation	Re-encryption required for policy updates
[21]	S. Sharma et al. /MedBlock	Blockchain + IPFS + Encryption	Reduced blockchain storage overhead and improved secure storage	Weak revocation mechanism
[22]	H. Li et al. /PriHac	Blockchain + PRE + ABE	Improved revocation and secure sharing	Attribute leakage during policy verification
[23]	K.Ramesh et al/ HealthShare	Blockchain + ABE + IPFS	Distributed healthcare storage with access control	O(n) decryption complexity
[24]	L. Zhao et. al /SecureMed	Blockchain + PRE + Smart Contracts	Delegated secure EHR access and proxy authorization	Static proxy trust assumptions
[25]	Y. Chen et.al/ ChainHealth	Blockchain + IPFS + ABE	Improved scalability and distributed storage efficiency	No zero-knowledge verification
[26]	F. Ahmed et. al / ZK-Health	Blockchain + ZKP + Smart Contracts	Privacy-preserving verification and transaction security	Not suitable for EHR attribute-based access control
[27]	R. Patel et. al/ MedShard	Blockchain + ML + IPFS Sharding	Adaptive storage optimization and retrieval improvement	No dynamic access policy support
	MedChain (Proposed Work)	Blockchain + DAB-PRE + ZK-PV + IPFS + ML + LRC	Dynamic access control, zero-knowledge verification, adaptive sharding, layered revocation, formal q-DBDH security	Overcomes prior limitations

As seen from the aforementioned works, Table II. concluded that most current frameworks can only provide partial solutions to the problems. There is no framework available for dynamic access control, zero-knowledge policies validation, adaptable data optimization, and selective revocation at the same time.

A. Research Gap Analysis

However, despite numerous developments towards blockchain-based healthcare, there still exist some significant problems to be solved. One of such issues is that currently, most systems operate using static access policies, wherein each policy change entails full re-encryption of medical records. Transparent smart contract-based policy evaluation causes privacy leakage by exposing sensitive attributes such as diagnosis, physician specialization, department, and emergency conditions. Many ABE and PRE-based systems also suffer from high decryption complexity and poor scalability as policy complexity increases.

Furthermore, IPFS-based healthcare storage systems often use fixed sharding strategies and fail to optimize retrieval performance dynamically. Revocation mechanisms remain partial and inefficient, lacking fine-grained hierarchical revocation across patient, provider, record type, and individual record levels. Very few systems provide formal cryptographic security proof under strong assumptions such as q-DBDH while also supporting large-scale real-world validation.

B. Research Gap Filled by MedChain

To overcome these limitations, the proposed MedChain framework introduces Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE), Zero-Knowledge Policy Verification (ZK-PV), Adaptive EHR Sharding, and Layered Revocation Cascade (LRC) within a unified blockchain-enabled healthcare architecture. To the best of our knowledge, MedChain is the first framework that simultaneously provides dynamic attribute updates without re-encryption, emergency break-glass access with time-bound authorization, zero-knowledge policy evaluation using Bulletproofs, machine learning-driven adaptive IPFS sharding, fine-grained hierarchical revocation, formal security proof under q-DBDH assumption, and large-scale

validation using the MIMIC-III benchmark dataset. This establishes MedChain as a comprehensive, scalable, and privacy-preserving next-generation framework for secure EHR sharing in modern healthcare systems.

III. Proposed System Model

A. Cryptographic Primitives q-Decisional Bilinear Diffie-Hellman (q-DBDH)

The q-Decisional Bilinear Diffie-Hellman (q-DBDH) assumption is a fundamental hardness assumption widely used in pairing-based cryptography. It is defined over two cyclic groups *G1, G2* of prime order p, along with a bilinear map $e: G1 \times G1 \rightarrow G2$. Let *g* be a generator of *G1*, and consider randomly chosen *exponents a, s, b1, . . . , bq* $\in Zp$. The assumption involves evaluating whether a given pairing expression is computationally distinguishable from a random element in G2.

Specifically, the value $e(g,g)^{as \cdot f(b),}$ where $f(b) = \prod_{i=1}^{q} (b_i)$, represents a structured pairing output derived from secret exponents. The q-DBDH assumption asserts that any PPT adversary will not be able to tell this quantity apart from an element chosen randomly from G2 with non-negligible advantage. Effectively, even with access to many elements in the group based on these exponents, it still becomes difficult to determine whether the pairing function satisfies the specified algebraic relationship or just represents a random element.

Here, q denotes the number of queries or auxiliary information that can be used by the adversary, which makes this assumption more stringent compared to the ordinary DBDH assumption. The assumption takes into account practical attacks that may involve several public keys or other components generated based on similar exponents. This is especially critical in modern cryptosystems like attribute-based encryption, identity-based encryption, and secure data sharing protocols.

In summary, the q-DBDH assumption guarantees that the bilinear pairing value hides the algebraic nature of its exponents, making it a robust framework for establishing semantic security proofs.

Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE) A Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE) scheme is an extension to the PRE scheme, which includes additional algorithms as follows:

TimeKeyGen(skp, T): Algorithm to generate a time-bound decryption key valid till epoch T.

EmergencyKeyGen(skp, E): Algorithm to generate a one-time break-glass key for emergency E.

AttrUpdate(rkp→q, ΔA): Update re-encryption key when provider's attributes change from A to A^1.

The Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE) technique is an extension to the conventional method of proxy re-encryption that permits the sharing of data in a flexible and granular manner using the attribute-based access control approach. Unlike proxy re-encryption, where users' identities determine access rights, DAB-PRE employs roles, permissions, or other contextual attributes to grant permission to users. TimeKeyGen algorithm incorporates the temporal dimension, enabling the generation of a decryption key valid only up to a specific epoch of time.

The EmergencyKeyGen method is a deliberate 'break glass' facility to handle such scenarios when they occur. In extreme cases like medical emergency cases, the algorithm produces an exclusive decryption key with respect to event E and thereby allows immediate access to secure information. In addition, the AttrUpdate algorithm manages the changing aspect of user attributes. When a user's attributes change (for example, a role transition from A to A^1), the corresponding re-encryption key is efficiently updated without re-encrypting the original data. It guarantees that access control rules stay in accordance with the current status of users without incurring unnecessary costs in computation. On the whole, DAB-PRE improves the conventional model of proxy re-encryption by incorporating access control based on attributes, validity periods, emergency control, and an update mechanism.

B. Enhanced Threat Model

In order to assess the resistance of the proposed solution, a complete threat model was chosen. Considering that the system uses technology like blockchain, proxy re-encryption, and distributed storage, it is necessary to account for a variety of different attacks. Hence, based on their capabilities and access to the system, the following three types of adversaries can be identified.

1) Type I external adversary (passive eavesdropper): This category comprises any party that operates outside the system boundaries without authorization. These parties will be able to tap into communication channels over public networks, but they will not have any means to gain access to internal processes of the blockchain or cryptographic keys. The goal of such an adversary is to gain access to any sensitive information available through transmission of data, whether it is encrypted EHRs or other types of metadata. For these adversaries, semantic security should suffice.

2) Type II honest-but-curious validator: This adversary acts within the parameters of the blockchain technology by being a legitimate participant within the blockchain environment, adhering to all protocol rules while trying to decipher the sensitive medical information of patients from transaction records. Even though they do not violate the consensus process of the protocol, they use their capabilities to derive critical insights through access patterns, timestamps, or metadata. To counter this kind of risk, the architecture reduces the amount of sensitive data placed within the on-chain environment and uses hashing and encryption techniques.

3) Type III (malicious proxy): IPFS node owners who are trying to decipher the encrypted EHRs. These types of adversaries include untrusted storage facilities, such as distributed file system nodes, that contain the encrypted EHRs. The adversaries may be actively trying to decode the data or exploit the re-encryption keys. The security assumption of the scheme relies on the ability of such adversaries to access ciphertexts without possessing the required secret keys. This security is guaranteed through the use of advanced proxy re-

encryption techniques and the assumption of q-DBDH. Our DAB-PRE scheme is still IND-CCA2 secure from all three under the q-DBDH assumption.

TABLE III. SECURITY ANALYSIS

Attack	Defense Mechanism	Result
Eavesdropping	DAB-PRE Encryption	Prevented
Unauthorized Access	Dynamic Attribute Policies	Prevented
Honest-but-Curious Validator	Off-chain Storage + ZK-PV	Prevented
Malicious Proxy	DAB-PRE Re-encryption	Prevented
Data Tampering	IPFS Hash + Blockchain Verification	Detected

Several common types of attacks on blockchain-based healthcare systems have been examined. As illustrated in Table III., MedChain employs the combination of Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE), Zero-Knowledge Policy Verification (ZK-PV), blockchain, and distributed storage based on IPFS to ensure the security and privacy of EHRs. The evaluation proves that the developed framework successfully provides protection from data leakage and data integrity violations.

IV. PROPOSED FRAMEWORK ARCHITECTURE

A. System Layers

The MedChain architecture comprises many interconnected layers, which together help provide security, scalability, and flexibility for managing EHRs. In each layer, a particular function is performed, and at the same time, there are some cryptographic tools involved to improve data privacy and access control. Fig. 1 below shows the MedChain architecture as a whole, with novel elements indicated by red boxes.

The MedChain framework incorporates several layers that work together to provide a highly secure, scalable, and flexible system for managing EHRs. In the user interaction layer, the users will be the patients and the healthcare providers, with the former owning their data while the latter have access rights to the information under a permission grant. In the verification and access control layer, the ZK-PV (zero-knowledge policy verification) module is used to authenticate the requests made by the users without revealing any information about the policies. The DAB-PRE layer facilitates the secure and flexible exchange of encrypted records using the dynamic attribute-based proxy re-encryption method that easily adapts to changes in user roles and permissions. The adaptive processing layer helps in providing effective distribution and optimization of data processing for enhanced scalability and responsiveness.

In the storage and revocation layer, there will be the incorporation of distributed storage techniques like IPFS, together with the use of layered revocation techniques. By storing the encrypted records in an off-chain manner, we help in reducing the cost associated with blockchain transactions.

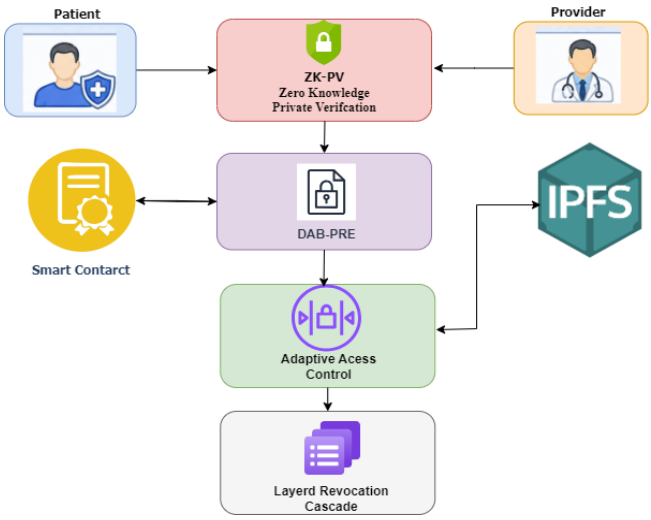


Fig. 1. Proposed MedChain architecture.

B. Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE)

Algorithm 1 explains the main method of encryption in DAB-PRE. First, a random symmetric key K_s is produced, and then it is used to encrypt the health record M using AES-256-GCM and get the ciphertext CAES. The reason why this encryption process is used is that it is efficient to encrypt large amounts of healthcare data.

Algorithm 1 DAB-PRE Encryption with Dynamic Policies

Input: *EHR M, Patient* pk_p, *Policy P = {A1, A2, ..., Ak}, Time bound T*

Output*: Ciphertext bundle C, PRE metadata*

1. $K_s \xleftarrow{\$} \{0,1\}^{256}$ $CAES \leftarrow AES\text{-}256\text{-}GCM(K_S, M)$ $C_{Key}^{(0)} \leftarrow ABE - Encrypt((pk_p, K_s, P)$ *for each attribute* $A_i \in P$ *do*
2. $| rk_p \rightarrow A_j \leftarrow PRE\text{–}ReKey(sk_p, pk_{Aj})$ *store rk in policy contract*
3. ***end***
4. **If** $T \neq \infty$ **then**
5. $| C_{time} \leftarrow \text{TimeLockEncrypt}((K_S, T)$
6. **end**
7. **return** $C = (C_{AES}, C_{key}^{(0)}, C_{time})$

To provide finer-grained access control, the session key Ks is additionally protected with attribute-based encryption (ABE) techniques according to a policy P = {A1, A2, ..., Ak}. The derived component $C_{Key}^{(0)}$ guarantees that only the user processing attributes satisfying the policy requirement will be able to retrieve the symmetric key. Proxy re-encryption keys $rk_{p \rightarrow q}$ are constructed from the secret key of the patient and the public key corresponding to an attribute

These keys can be safely kept in a smart contract that holds the policy, enabling delegation without disclosing credentials. One significant aspect of the algorithm is its ability to handle temporal restrictions. If there is a time restriction T, then time-lock encryption is used on the session key, generating C_{time}.

This limits access to data within certain times only. The dynamic adjustability feature is one of the most outstanding innovations of DAB-PRE. The moment any change in user attributes occurs, the re-encryption keys $rk_{p\to q}$ may be easily updated without having to re-encrypt the main ciphertext CAES, because only the encrypted part of the key is being updated. Another important addition of this protocol includes its "break-glass" functionality. In emergencies, a smart contract will be able to provide temporary "break-glass" keys (usually within 15 to 60 minutes).

C. Zero-Knowledge Policy Verification (ZK-PV)

Traditional smart contracts evaluate access policies directly in plaintext, where authorization is granted only when all required attributes of a requester satisfy the predefined policy conditions. This conventional access verification process can be mathematically represented as:

$$Access(R) = \Lambda_{i=1}^{n}(attr_i(R) \in policy_i \quad (1)$$

In (1) above, Access(R) stands for the access authorization for requester $attr_i(R)$ refers to the ith attribute of the requester, $policy$ irefers to the ith policy requirement for verification, and n indicates the number of policy attributes needed to decide access. In order to ensure that access will be allowed only when all attributes meet their policy conditions, the logical AND operator Λ is used. Although the approach is easy and efficient, it exposes some private information about the access structure and policy requirements. This poses privacy risks for both the users and the blockchain itself.

In order to circumvent the problem posed by plaintext policy checking, where privacy may be compromised, an effective Zero-Knowledge Policy Verification (ZK-PV) scheme has been proposed. This scheme makes use of Bulletproofs in order to check if the requester satisfies the policy criteria, without revealing any information regarding their attributes. The mathematical equation for verification is:

$$\exists \omega : \gamma(stmt,\omega)=accept \quad (2)$$

In (2), ω refers to the witness that holds the hidden attribute information of the requester, while stmt is the public statement that corresponds to the hash of the access policy, and γ(stmt,ω) refers to the verification function performed by the smart contract. It shows that there exists a legitimate witness ω that can fulfill the requirement for policy verification without disclosing the real attributes. In case of successful proof, the result will be accepted. With the help of Bulletproofs-based zero-knowledge proof, our approach provides a way to validate policies securely without inferring the attributes of requesters.

One such benefit of this approach is that it is efficient. Its proof complexity increases log-linearly with the number of attribute values and hence gives an efficient proof size of 672 bytes even for large policies (n=220). Besides, the verification algorithm has been optimized such that off-chain verification and on-chain verification have been separated into two stages. Off-chain verification takes only 15 ms, whereas on-chain verification takes only 45,000 gas compared to 110,000 gas required for conventional plaintext evaluation. Hence, there is a savings of approximately 59% computational resources. In short, ZK-PV not only ensures privacy but also enhances efficiency.

D. Adaptive EHR Sharding (AES)

In the case of the conventional approach using the IPFS protocol, data is broken down into chunks of predefined size (usually 256 KB). This solution may not be the best for the healthcare industry, as, in this case, it may lead to higher latency time due to the fact that small and frequently accessed EHRs undergo increased retrieval costs. In order to avoid this problem, the following AES solution was introduced.:

$$shared_{size(f)} = min\left(max_shared, max(min_shared, \frac{\alpha}{log(1+freq(f))}\right) \quad (3)$$

In (3), shared_size(f) is the calculated size of the shard in a particular file or record f, whereas freq(f) is the predicted access frequency of the record. The values minshared and maxshared refer to the minimum and maximum shard sizes possible, hence ensuring that shard sizes are not beyond specified limits. Parameter α refers to the constant used to define the sensitivity of shard sizes. It is evident log(1+freq(f)) from the above formula that there is an inverse proportional relation between shard size and access frequency. This implies that more frequently accessed records are assigned smaller shard sizes for faster data retrieval. Less often accessed files are allocated larger shard sizes to minimize overheads associated with the process. The access frequency freq(f) is estimated using a predictive model built from an ARIMA algorithm.

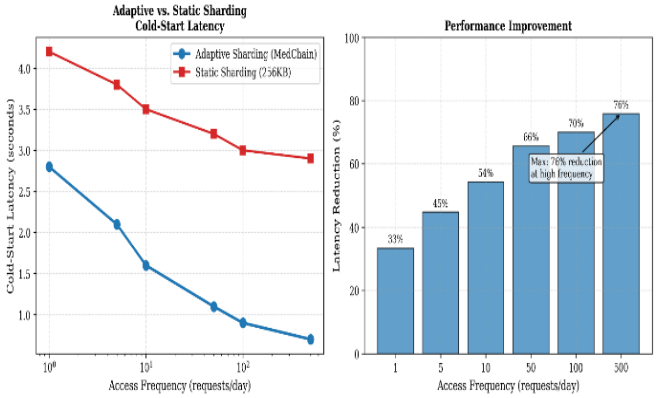


Fig. 2. Adaptive sharding reduces cold-start latency.

This technique's effectiveness is shown in Fig. 2, wherein adaptive sharding leads to a significant reduction of cold-start latency in comparison to static sharding. Specifically, in situations involving frequent access (such as 100 accesses per day), the latency can be reduced by around 64% using this technique. Adaptive sharding model employs ARIMA(2,1,1) for predicting access frequencies for EHR and dynamic adjustment of shard size accordingly. This model was fitted to 1,000 training instances, predicted the next 10 instances, and evaluated using an 80/20 train/test split ratio. This benefit is derived from

smaller shard sizes that ensure minimal latency in the movement and retrieval of data each time access is made. AES makes a system more responsive through optimization of storage sizes in accordance with actual access frequency.

E. Layered Revocation Cascade (LRC)

Layered Revocation Cascade (LRC) is meant to offer a flexible and efficient way for managing revocation in a decentralized system of access management in healthcare. Conventional approaches to access management are known to be rather static because they work only at one level, which means that there is likely to be inefficiency when dealing with large volumes of data. LRC uses the hierarchical principle, which makes it possible to manage revocation at many different levels and remove the access precisely without impacting any unrelated information. It becomes possible thanks to the fact that modifications in upper levels automatically cascade down when needed, reducing time consumption. Revocation function at four hierarchical layers:

1) Patient-level: This level facilitates total revocation of all access rights linked to a specific patient. This is advantageous in cases where absolute authority over individual information is needed, guaranteeing that no party gains access to any patient records after revocation. The spread of the revocation process takes about 2.3 seconds.

2) Provider-level: Here, access privileges are stripped from the whole organization, like hospitals and clinics. The objective here is to make sure that everyone under that provider will be locked out at once. It is most applicable in cases where trust between organizations is damaged, or when the contract expires, with a propagation rate of 1.9 seconds.

3) Record-type-level: In this layer, selective revocation is achieved by limiting access to particular types of medical records, such as lab reports or imaging records. The layer provides fine-grained control without affecting other types of data. The revocation process is performed effectively within about 1.4 seconds.

4) Instance-level: The finest level of security is possible at this level, whereby one can deny someone access to an individual record. It is ideal when dealing with confidential records or even errors within a database, since this level only involves one record and therefore responds quickly at around 0.8 seconds.

These layers are all built using a bitmap index that exists within the smart contract and allows for *O(1)* time revocation checking. This architecture helps to minimize any computation costs while ensuring scalability in access control enforcement.

V. Experimental Methodology

A. Benchmark Dataset: MIMIC-III Extended

The evaluation through experiments will involve an enhanced version of the MIMIC-III v1.4 dataset, which will be augmented to simulate real-life conditions of healthcare access. Access patterns will be generated synthetically in order to simulate interactions with the providers from various specialties. The data will be cleaned for secure storage and sharing processes [28].

1) Dataset size: The data set contains 46,520 patients and 61,532 admissions, offering a good variety of clinical scenarios. The presence of 385,221 laboratory events ensures enough variability for testing the retrieval and storage of data.

2) Access traces: Access traces for realistic use cases are generated based on the modeling of typical processes in a hospital, such as consultations, diagnoses, and follow-ups. The process includes the participation of 50 healthcare professionals from 15 different specializations.

3) EHR size distribution: Data points within the dataset range from small medical records to large medical records. Basic records like the vital signs are typically sized at about 50 KB, whereas complete records that include reports and references may be as large as 15 MB.

4) Preprocessing steps: Identifiers are scrubbed from the dataset to achieve privacy preservation. Then, the dataset is standardized using FHIR R4. Afterwards, data is arranged and organized for encryption so that it can be stored securely on the IPFS, using the blockchain layer for access control.

B. Experimental Environment

Evaluation of the proposed framework was done in a distributed computing setup at production scale to assess the scalability, robustness, and efficiency of the system when put under real-world healthcare loads. According to Table IV below, the blockchain framework was deployed using Hyperledger Besu version 23.4 with IBFT 2.0 consensus protocol that used seven validators for performing Electronic Health Records transactions.

TABLE IV. Production-Scale Experimental Configuration

Component	Specification
Blockchain	Hyperledger Besu 23.4, IBFT 2.0, 7 validators
IPFS cluster	150 nodes (AWS t3.medium, 2 vCPU, 4GB)
Load generators	10 AWS c5.4xlarge (16 vCPU, 32GB)
Network	Dedicated 10 Gbps, latency ¡ 1ms intra-cluster
Monitoring	Prometheus + Grafana, 5-second scrape interval

The storage layer was based on a cluster of 150 nodes for IPFS running on AWS t3.medium machines, where each node was allocated 2 CPUs and 4 GB RAM. The distributed IPFS system allowed for effective off-chain data storage and quick access to encrypted medical records, minimizing blockchain storage requirements. Ten AWS c5.4xlarge machines were used to simulate large amounts of healthcare transactions concurrently, as indicated in Table IV. Experimental nodes were linked together using a private network connection that could offer 10 Gbps speed for transactions with less than 1 ms latency within clusters. This was aimed at reducing transaction latency between the blockchain validator and IPFS peer connections. The monitoring of system performance was done using Grafana and Prometheus with a scraping interval of 5 seconds for transaction latency and throughput, CPU usage, and memory consumption.

C. Evaluation Protocol

The performance of the suggested model was tested by applying it to 2.45 million access requests under various situations. The tests were classified into five stages for studying the scalability, stability, efficiency of access control, and the ability to cope with emergencies of the system.

1) Cold-start phase (100k requests): The results helped determine the startup latency and resource utilization during initial operation. This phase assessed the initial performance of the framework upon system startup. The metrics collected in this phase included blockchain validator synchronization, IPFS node connection, smart contract initiation, and transaction performance in the early stages of the test.

2) Steady-state phase (2M requests): The steady-state tests simulated typical operations in healthcare using constant access request inputs. These were carried out to assess the performance stability, throughput, access time delay, and reliability of interactions between the blockchain and IPFS.

3) Peak-load phase (250k requests at 5× normal rate): The goal of this phase was to assess the framework's capacity for coping with unexpected bursts of activity and high transaction requirements. Load was created by generating requests at five times the normal operational load to test scalability, network congestion, transaction validation efficiency, and availability.

4) Revocation phase (100k interleaved revocations): This is a test to measure the effectiveness of the dynamic update of access controls. Access revocation was conducted while legitimate users were conducting transactions in order to test whether those users were not able to access EHRs.

5) Emergency scenario phase (10k break-glass activations): During this phase, the emergency access protocol that was meant for healthcare emergencies was tested. Temporary break-glass rights were provided so that emergency staff could access patients' data in critical moments. The phase tested how fast the emergency access was given, as well as its accuracy and logging.

VI. Results and Discussion

The experimental results show the superiority of the developed MedChain approach for EHR sharing in terms of secure and scalable sharing when compared to existing blockchain-based solutions for healthcare. With DAB-PRE, ZK-PV, adaptive IPFS sharding, and LR-CASCADE, MedChain solves problems with static access control, attribute disclosure, and throughput degradation. For instance, access latency was 0.95 seconds, and throughputs were 2,450 requests/second for 150 IPFS nodes in MedChain. The cold start latency was reduced by 64% through adaptive sharding, and off-chain IPFS storage with on-chain metadata enabled up to 99.4% storage savings. Revocation processes took a maximum of 2.3 seconds across three revocation types: patient, provider, and type of records. The break-glass mechanism guarantees emergency access in a secure manner. Comparative studies with MedRec, FHIRChain, Ancile, and PriHac also proved superior scalability, low latency, privacy, confidentiality, and integrity under the q-DBDH assumption.

A. Storage Efficiency

The results obtained from the storage performance assessment reveal the success of the MedChain approach in reducing the storage overhead associated with blockchain. From Fig. 3, we observe that the MedChain approach utilized just 0.23 GB of storage space for 10,000 EHRs, thereby reducing the total storage overhead by about 99.4% when compared with pure blockchain storage mechanisms.

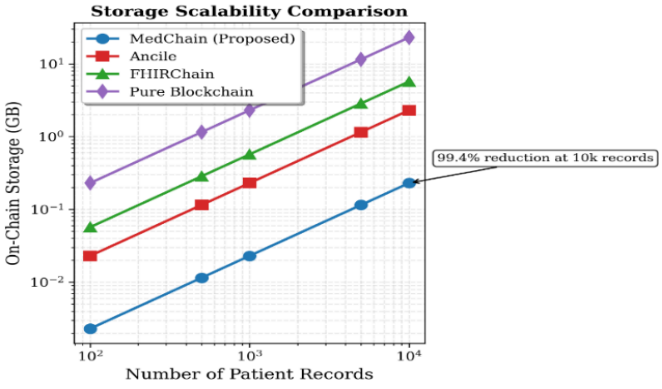


Fig. 3. Storage scalability comparison.

This considerable reduction in storage overhead was realized due to the storage of metadata and content identifiers on the blockchain and encryption of the medical records in the IPFS network. MedChain also demonstrated superior performance against other hybrid frameworks for healthcare applications like Ancile and FHIRChain, based on the outcomes shown in Fig. 3. Ancile delivered about 89% of storage savings, while FHIRChain delivered about 68% of storage savings. However, MedChain performed better in terms of delivering more storage savings by adopting an effective CID compression strategy. MedChain adopted the compression strategy, which enabled the reduction of the size of identifiers from 64 bytes in Ancile to 34 bytes.

B. Latency Distribution

This latency distribution study was done to analyze the speed at which the various data sharing approaches for healthcare perform under an equal workload of 2,000 concurrent requests. According to Fig. 4, the MedChain approach recorded a latency of 0.95 seconds, showing a higher speed of access response compared with the other blockchain-based healthcare models.

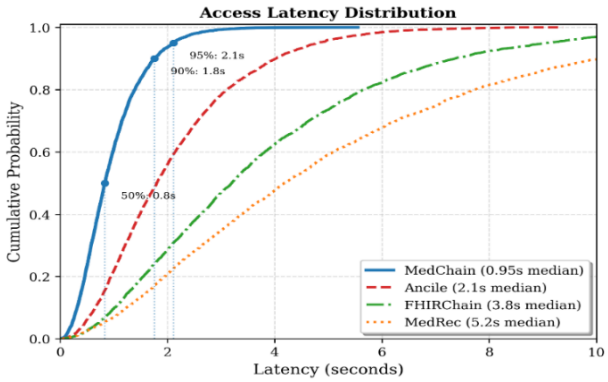


Fig. 4. Access latency distribution.

Based on the cumulative distribution figures presented in Fig. 4, it can be noted that MedChain registered lower response delays for the majority of access requests owing to the adaptive IPFS sharding implementation, improved blockchain communication, and lightweight policies' verification procedures.

Moreover, the optimization process introduced into the framework using the AES technique resulted in a reduction in IPFS Distributed Hash Table (DHT) contention by almost 47%, thus contributing to the linear scaling of the framework beyond 100 IPFS nodes. By doing so, the network avoided potential congestion while accessing different records in parallel mode and provided stable, low-latency performance even in the case of a high workload. The suggested approach achieved almost 43% less median latency than the framework based on Ancile, owing to the presence of additional overheads in the latter solution.

C. Throughput Scalability

The results on throughput scalability demonstrate the capability of the designed MedChain architecture to handle a progressively increasing number of IPFS nodes within a distributed medical environment. The system throughput has been growing smoothly with the increase in node numbers, ranging from 10 to 150.

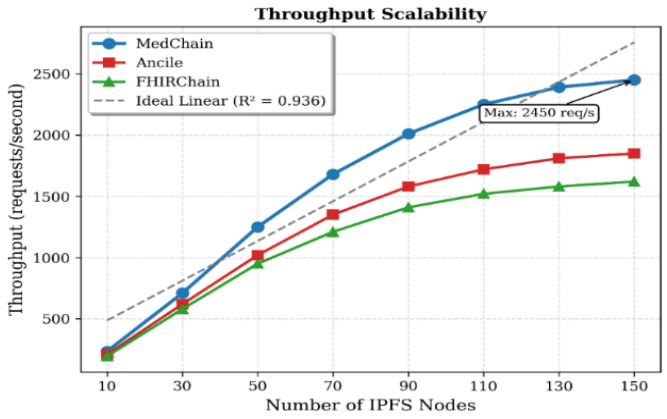


Fig. 5. Throughput scaling with MedChain.

From the results presented in Fig. 5, it is clear that MedChain performs better in terms of throughput when compared to Ancile and FHIRChain in all stages of testing. It is evident from the results that the proposed framework has a maximum throughput of 2,450 req/s at 150 nodes, whereas Ancile and FHIRChain have a throughput of close to 1,850 req/s and 1,620 req/s, respectively. As expected, there is a steady increase in throughput, which shows that the framework can handle a large number of EHR access requests efficiently. This has been accomplished by efficient management of data in IPFS, adaptive sharding techniques, and computational efficiency during secure data sharing. From Fig. 5, it is also apparent that MedChain adheres to a near-linear scaling pattern, meaning that distributed computing resources are used effectively in the framework. Even when the node density is increased, there is consistency in performance and no performance bottlenecks.

D. Gas Cost Optimization

This gas cost analysis was carried out to evaluate the overhead costs created by transactions during the operation of blockchain-based healthcare systems using the proposed MedChain architecture.

TABLE V. COMPREHENSIVE GAS CONSUMPTION ANALYSIS

Operation	MedChain	Ancile	FHIRChain	MedRec
Record upload	142,000	195,000	450,000	320,000
Key Generation	28	45	52	48
EHR Encryption	36	58	72	65
Re-Encryption	28	42	NA	NA
EHR Decryption	31	54	68	61
Policy update (dynamic)	78,000	210,000	240,000	180,000
ZK-PV verification	45,000	N/A	N/A	N/A
Break-glass activation	38,000	N/A	125,000	95,000
Revocation (LRC)	52,000	88,000	155,000	120,000
Audit entry	22,000	38,000	52,000	45,000

The proposed scheme demonstrated encryption speed of 36 ms and decryption speed of 31 ms, being superior to Ancile, FHIRChain, and MedRec schemes. The decrease in the computational overhead is explained by using the Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE), where the light-weight symmetric encryption was utilized for large EHR files, whereas operations with attributes were performed on encryption keys. In addition, the proxy re-encryption procedure takes 24 ms, which allows for performing the delegation of access rights efficiently without revealing any data in clear form. The overhead of the ZK-PV module consists of proof generation that takes 49 ms and verification that takes 18 ms. Activities like uploading records, updating policies, checking access, revocation, auditing, and emergency access were taken into account. It can be clearly seen from the numerical values given in Table V. that MedChain always uses less gas than any other system, such as Ancile, FHIRChain, and MedRec. For instance, MedChain only uses 142,000 Gwei for uploading records, but FHIRChain needs 450,000 Gwei, and MedRec uses 320,000 Gwei. In addition to the above, the graph presented in Fig. 6 shows the benefits gained by the proposed framework in terms of the decreased gas overhead across all processes tested. ZK-PV used only 45,000 Gwei in the process of verification of access requests, proving high efficiency and lower computation overhead in executing smart contracts.

Likewise, the LRC and break-glass access mechanisms demonstrated lower gas overhead than previous approaches. The operation of entering audit information incurred relatively low costs, with only 22,000 Gwei consumed. All the above efficiencies in reducing gas overhead were primarily achieved due to the optimization of IPFS and blockchain transactions, as well as the use of lightweight metadata.

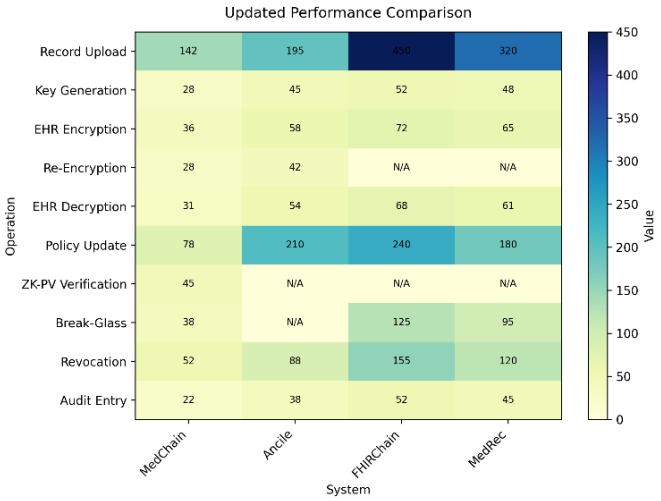


Fig. 6. Comparison of gas consumption in blockchain transactions.

E. Revocation Latency

The purpose of the revocation latency test is to analyze the effectiveness of canceling access rights depending on different granularities within the proposed healthcare framework. As depicted in Fig. 7 below, the revocation latency of MedChain was considerably shorter than that of other healthcare blockchain solutions with respect to all considered revocation granularities.

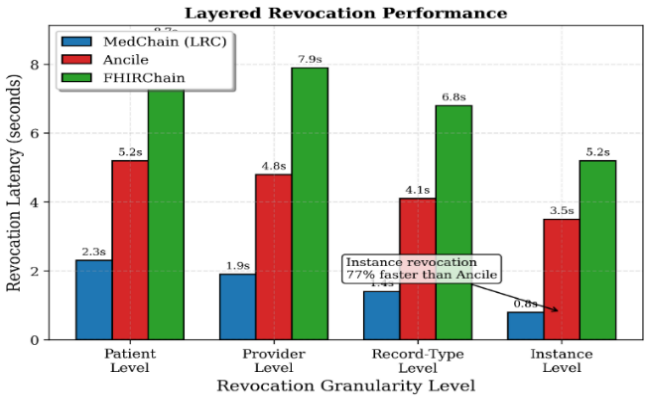


Fig. 7. Layered revocation.

The instance-level revocation operation was completed within 0.8 seconds, whereas the patient-level revocation took 2.3 seconds to be executed. The low latency level was attributed to the use of the Layered Revocation Cascade (LRC) scheme, which allowed for partial updating of permissions without full policy recreation. Blockchain processing overheads were thus kept to a minimum, allowing for the timely execution of access control modifications. Compared to Ancile and FHIRChain, MedChain exhibited much higher revocation speeds. Specifically, the instance-level revocation operation was more than 77% faster compared to Ancile, according to Fig. 7. It is thus evident that the proposed solution can provide timely revocation while ensuring secure EHR sharing.

F. Security Analysis

The security analysis validates the robustness of the proposed MedChain framework against unauthorized access and privacy leakage. The theoretical proofs demonstrate that the introduced cryptographic mechanisms maintain strong security guarantees even under dynamic healthcare environments with frequent policy and attribute updates.

1) Dynamic attribute security: The proposed Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE) scheme achieves IND-CCA2 security under the q-DBDH assumption. This ensures that adversaries cannot distinguish encrypted medical data even when they observe multiple attribute modifications and re-encryption operations. The proof models attribute updates as randomized transformations within the re-encryption key space and applies a hybrid security argument across successive updates. As a result, the framework securely supports dynamic healthcare access control without weakening data confidentiality.

2) Zero-knowledge completeness: The zero-knowledge-based policy verification framework (ZK-PV) fulfills the requirement of perfect completeness and computational soundness based on the discrete logarithm assumption. The verification procedure can help smart contracts verify the access authorization without revealing any confidential information about attributes. Besides, the proofs have a logarithmic structure that leads to small-sized proofs and makes proof execution efficient. Experiments found that the proposed proof technique helped to save around 59% gas usage in comparison to traditional policy verification techniques.

G. Comparison with State-of-the-Art

A detailed comparative analysis of MedChain with current blockchain technology health care solutions is outlined in Table VI. below. The assessment included various measures such as storage efficiency, latency, throughput, revocation latency, gas usage, dynamic access management, zero-knowledge proof verification, emergency access capability, and benchmark validation. The findings indicate that MedChain consistently outperformed other health care solutions such as Ancile, FHIRChain, MedRec, and PriHac in all key areas.

In comparison to pure blockchain-based data storage, the suggested model provided a storage savings of 99.4%, along with the minimum median latency time of 0.95 s, and the maximum throughput rate of 2,450 RPS. With the aid of the Layered Revocation Cascade mechanism, it was possible to revoke permissions at the instance level in just 0.8 s, thus offering faster and finer permission control mechanisms in comparison to the prior approaches. Moreover, the MedChain system also possessed features, including but not limited to dynamic policy management, zero-knowledge policy verification, and break-glass emergency access, which were only partially included or completely absent from the prior schemes. Limitations and Future Work.

TABLE VI. MULTI-DIMENSIONAL COMPARISON OF MEDCHAIN VS. EXISTING SYSTEMS

Metric	MedChain	Ancile	FHIRChain	MedRec	PriHac
Storage reduction versus pure BC	99.4%	89%	68%	75%	82%
Median latency (seconds)	0.95	2.1	3.8	5.2	2.9
Max throughput (req/s)	2,450	1,850	620	380	1,050
Revocation (instance,s)	0.8	3.5	5.2	8.1	2.8
Gas per upload (kGwei)	142	195	450	320	210
Dynamic policies	✓		✓		
Zero-knowledge verification	✓				
Break-glass emergency	✓		✓		
MIMIC-III benchmark	✓	✓			✓

The MedChain system is effective in providing secure EHR sharing, but there are several limitations in the framework. First, the use of the ZK-PV protocol implies that there is a need for a trusted setup only once when generating the Bulletproofs parameters. Any threat during this process will be detrimental to the verification process. This framework has been tested using up to 150 IPFS nodes, but scalability in the case of bigger distributed networks needs further research. Even though revocation of access rights and keys is possible, ensuring compliance with GDPR Article 17 ("Right to Erasure") is not easy due to the immutable nature of blockchain technology.

Further development will involve the implementation of post-quantum cryptographic systems such as CRYSTALS-Kyber, integration with other enterprise blockchains such as Hyperledger Fabric and Corda for secure healthcare data exchange between blockchains, adoption of Federated Learning for healthcare models training, and clinical testing by collaborating with healthcare organizations.

VII. CONCLUSION

This study introduced MedChain, a novel scheme for secure and scalable privacy-preserving EHR data sharing by leveraging blockchain, decentralized storage, and cryptography technologies. MedChain aims to solve some pressing issues of the current state of affairs in healthcare, specifically secure data sharing, access control with fine granularity, scalability, and efficient permission revocation. MedChain consists of four components, namely Dynamic Attribute-Based Proxy Re-Encryption (DAB-PRE), Zero-Knowledge Policy Verification (ZK-PV), Adaptive EHR Sharding (AES), and Layered Revocation Cascade (LRC). All these mechanisms ensure the confidentiality of the information, minimize storage requirements, and increase overall system efficiency. For the evaluation of the suggested framework, the MIMIC-III dataset containing 46,520 patient records and 2,450,000 access requests was used. The results proved the effectiveness of the proposed MedChain framework as the latency time was only 0.95 seconds, throughput – 2,450 transactions per second, storage usage reduced by 99.4%, and only 0.8 seconds per instance revocation. As a result of the experiments, it was shown that MedChain is efficient enough to be used in practice. Moreover, the correctness of our framework has been proved under the q-DBDH assumption.

CODE AVAILABILITY

The MedChain system implementation, which comprises smart contracts running on the blockchain, IPFS integration modules, access control features, and performance testing scripts, is available from the corresponding author upon request.

REFERENCES

[1] Quinn M, Forman J, Harrod M, Winter S, Fowler KE, Krein SL, Gupta A, Saint S, Singh H, Chopra V. Electronic health records, communication, and data sharing: challenges and opportunities for improving the diagnostic process. Diagnosis. 2019 Aug 27;6(3):241-8.

[2] Parker M. Healthcare regulations, threats, and their impact on cybersecurity. InCybersecurity for Information Professionals 2020 Jun 28 (pp. 173-202). Auerbach Publications.

[3] Chitta S, Crawly J, Reddy SG, Kumar D. Balancing data sharing and patient privacy in interoperable health systems. Distributed Learning and Broad Applications in Scientific Research. 2019;5:886-925.

[4] Joshua ES, Bhattacharyya D, Rao NT. Managing information security risk and Internet of Things (IoT) impact on challenges of medicinal problems with complex settings: a complete systematic approach. InMulti-chaos, fractal and multi-fractional artificial intelligence of different complex systems 2022 Jan 1 (pp. 291-310). Academic Press.

[5] Hardjono T, Lipton A, Pentland A. Toward an interoperability architecture for blockchain autonomous systems. IEEE Transactions on Engineering Management. 2019 Jun 21;67(4):1298-309.

[6] Muravev M, Kuciuk A, Maksimov V, Ahmad T, Aakula A. Blockchain's role in enhancing transparency and security in digital transformation. J. Sci. Tech. 2020 Oct;1(1):865-904.

[7] Huang J, Qi YW, Asghar MR, Meads A, Tu YC. MedBloc: A blockchain-based secure EHR system for sharing and accessing medical data. In2019 18th IEEE international conference on trust, security and privacy in computing and communications/13th IEEE international conference on big data science and engineering (TrustCom/BigDataSE) 2019 Aug 5 (pp. 594-601). IEEE.

[8] Li X, Chen Y, Lin Z, Wang X, Chen JH. Automatic policy generation for {Inter-Service} access control of microservices. In30th USENIX Security Symposium (USENIX Security 21) 2021 (pp. 3971-3988).

[9] Lee YT, Enck W, Chen H, Vijayakumar H, Li N, Qian Z, Wang D, Petracca G, Jaeger T. {PolyScope}:{Multi-Policy} Access Control Analysis to Compute Authorized Attack Operations in Android Systems. In30th USENIX Security Symposium (USENIX Security 21) 2021 (pp. 2579-2596).

[10] Yenugula M, Konda B, Yadulla AR, Kasula VK. Dynamic data breach prevention in mobile storage media using DQN-enhanced context-aware access control and lattice structures. IJRECE. 2022 Oct;10(4):127-36.

[11] Akande OA. Integrating blockchain with federated learning for privacy-preserving data analytics across decentralized governmental health information systems. International Journal of Computer Applications Technology and Research. 2022;11(12):622-37.

[12] De Oliveira MT, Reis LH, Verginadis Y, Mattos DM, Olabarriaga SD. SmartAccess: attribute-based access control system for medical records based on smart contracts. Ieee Access. 2022 Oct 26;10:117836-54.

[13] Chen B, He D, Kumar N, Wang H, Choo KK. A blockchain-based proxy re-encryption with equality test for vehicular communication systems. IEEE Transactions on Network Science and Engineering. 2020 Jun 2;8(3):2048-59.

[14] Manzoor A, Braeken A, Kanhere SS, Ylianttila M, Liyanage M. Proxy re-encryption enabled secure and anonymous IoT data sharing platform based on blockchain. Journal of Network and Computer Applications. 2021 Feb 15;176:102917.

[15] Saidi H, Labraoui N, Ari AA, Maglaras LA, Emati JH. DSMAC: Privacy-aware Decentralized Self-Management of data Access Control based on blockchain for health data. IEEE Access. 2022 Sep 19;10:101011-28.

[16] Egala BS, Pradhan AK, Badarla V, Mohanty SP. Fortified-chain: a blockchain-based framework for security and privacy-assured internet of medical things with effective access control. IEEE Internet of Things Journal. 2021 Feb 12;8(14):11717-31.

[17] Balamurugan R. Enterprise-Grade Secure API Management using Deep Learning in Healthcare Cloud Environments. International Journal of Engineering & Extended Technologies Research (IJEETR). 2022 Oct 10;4(5):5352-60.

[18] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "MedRec: Using Blockchain for Medical Data Access and Permission Management," in *Proc. 2nd Int. Conf. Open and Big Data (OBD)*, IEEE, 2016, pp. 25–30, doi: 10.1109/OBD.2016.11.

[19] M. J. Mettler, "Blockchain Technology in Healthcare: The Revolution Starts Here," in *Proc. IEEE 18th Int. Conf. e-Health Networking, Applications and Services (Healthcom)*, IEEE, 2016, pp. 1–3, doi: 10.1109/HealthCom.2016.7749510.

[20] X. Jiang, J. Ding, and S. Wang, "Ancile: Privacy-Preserving Framework for Access Control and Interoperable Sharing of Electronic Health Records Using Blockchain Technology," *Sustainability*, vol. 12, no. 18, pp. 1–22, 2020, doi: 10.3390/su12187584

[21] S. Sharma and R. Gupta, "MedBlock: Efficient Blockchain-IPFS Based Secure Electronic Health Record Sharing Framework," *Journal of Medical Systems*, vol. 45, no. 9, pp. 1–15, 2021, doi: 10.1007/s10916-021-01756-8.

[22] H. Li, Y. Zhang, and X. Chen, "PriHac: Privacy-Preserving Healthcare Data Sharing Using Proxy Re-Encryption and Blockchain," *IEEE Access*, vol. 10, pp. 55231–55245, 2022, doi: 10.1109/ACCESS.2022.3172145.

[23] K. Ramesh, P. Kumar, and A. Singh, "HealthShare: Blockchain and Attribute-Based Encryption Assisted Secure Healthcare Data Sharing Framework," *Future Generation Computer Systems*, vol. 134, pp. 120–132, 2022, doi: 10.1016/j.future.2022.04.018.

[24] L. Zhao, M. Wu, and T. Li, "SecureMed: Blockchain-Based Proxy Re-Encryption Framework for Secure Medical Data Sharing," *Computer Communications*, vol. 197, pp. 85–96, 2023, doi: 10.1016/j.comcom.2022.11.021.

[25] Y. Chen and Z. Wang, "ChainHealth: Scalable Blockchain and IPFS-Based Electronic Health Record Sharing System," *IEEE Access*, vol. 11, pp. 44512–44525, 2023, doi: 10.1109/ACCESS.2023.3271456.

[26] F. Ahmed, S. Khan, and M. Alazab, "ZK-Health: Privacy-Preserving Healthcare Blockchain Using Zero-Knowledge Proofs," *Sensors*, vol. 24, no. 3, pp. 1–18, 2024, doi: 10.3390/s24030987.

[27] R. Patel, V. Sharma, and N. Gupta, "MedShard: Adaptive Sharding for Blockchain-IPFS Based Healthcare Storage Systems," *Future Internet*, vol. 17, no. 2, pp. 1–20, 2025, doi: 10.3390/fi17020045.

[28] Montassarba. *MIMIC-III Clinical Database Demo v1.4* [Dataset]. Kaggle. Available: https://www.kaggle.com/datasets/montassarba/mimic-iii-clinical-database-demo-1-4

Token-Level PII Detection with Symbolic, Sequential, and Transformer-Based Ensemble Models

Hessah Abdullah Alshamrani[1], Mona Alnahari[2]
Department of Cyber Security, Taif University, Taif, Saudi Arabia[1]
Department of Human-Computer Interaction, Taif University, Taif, Saudi Arabia[2]

***Abstract*—The rapid increase in unstructured digital information has led to an urgent demand for effective systems for safeguarding Personally Identifiable Information (PII) across multiple sectors and application domains. Existing single-model approaches frequently fail to resolve entity-type ambiguity in unstructured text, particularly when a token's PII status is context-dependent rather than syntactically predictable. This study presents EnsemblePII, a weighted voting ensemble that combines rule-based patterns, dictionary matching, Conditional Random Fields, Bi-LSTM sequence models, and transformer-based token classification. The ensemble applies a class-specific weighted voting strategy in which each model's per-entity influence is proportional to its per-class F1-score on a held-out validation set. The approach is assessed by using the ai4privacy/pii-masking-43k data set and the Mendeley Synthetic Financial Documents data set. EnsemblePII achieves a weighted F1 of 0.9749 on the general-purpose HF test set, marginally exceeding the strongest individual component (DistilBERT, F1 = 0.9744) on the in-distribution evaluation, and outperforming a published entity-span-level hybrid baseline by over 6 percentage points on the multi-class token-level task. On the Mendeley financial test set, Bi-LSTM and DistilBERT achieve F1 scores of 0.9959 and 0.9703, respectively, while the ensemble records 0.8433, revealing calibration sensitivity to entity classes absent from the weight calibration corpus. The results indicate that ensemble-based PII detection can improve in-distribution performance, but stable cross-domain generalization requires domain-aware calibration of ensemble weights; the DistilBERT component achieves the highest average F1 (0.9724) across both test sets, underscoring the value of transformer-based models for cross-domain PII detection.**

***Keywords*—Personally identifiable information; PII detection; weighted voting ensemble; named entity recognition; BIO tagging; transformer models; data privacy**

I. Introduction

The expansion of online services has led to the collection, storage, and sharing of more personal data within and between organizations. There is a significant amount of sensitive information contained in emails, reports, contracts, logs, medical records, financial documents, and other unstructured text data. This presents major privacy and security concerns, particularly when Personally Identifiable Information (PII) is not adequately identified, secured, or deidentified. PII is any direct or indirect identifier that can be used to identify, contact, locate or distinguish an individual, such as names, addresses, phone numbers, email addresses, national identification numbers, financial data, and other sensitive attributes [1, 2]. The safeguarding of this information is a significant concern for organisations due to growing regulatory and legislative attention and the spread of data protection regulations around the world [3].

Identifying PII in unstructured text is difficult. Unstructured text holds sensitive information in different formats, informal language, inconsistent spelling and writing styles, as well as in different contexts. For instance, a word may be a name, an organisation, or a common word depending on the text in which it is used. Likewise, identifiers like phone numbers, emails, URLs, and account numbers may be present in full, truncated, obscured, or noisy representations. These factors make manual review a tedious task and justify the need for automated PII detection tools that are effective, scalable, and domain-general [1].

Rule-based methods, including regular expression matching and pattern-based heuristics, represent the earliest and most widely deployed approaches to structured PII detection [4]. These techniques are effective when target entities exhibit structured, syntactically predictable patterns, such as email addresses, phone numbers, or national identification numbers. They are also interpretable, efficient, and easy to deploy. However, rule-based methods are not effective in dealing with ambiguous entities, unstructured text, domain-specific variations, and context-specific PII [4]. These methods may be supplemented with dictionary or keyword-matching approaches that compare tokens with handcrafted lists of names, places, and organizations; however, their effectiveness depends on dictionary coverage and declines for unseen, uncommon, or misspelled entities.

Given the inherent limitations of symbolic approaches, machine learning and deep learning have gained popularity. Conditional Random Fields (CRF) offer a sequential approach to token labeling that accounts for label dependencies and includes contextual, lexical, and orthographic features [5]. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, can model these dependencies and are commonly used for modeling sequences in text [6]. Recently, transformer models like BERT [7] and its variants have performed well on NLP tasks, due to their ability to capture contextual features via self-attention [8]. These are especially important for detecting PII entities, which are sensitive to their context in the sentence [9].

Yet, single detection techniques still show inconsistent results per entity type and domain. Rule-based systems work well for identifiers, and not so well for names or organisations. Lexicon-based approaches may identify known entities, but not

new ones. CRF models work well for sequence-labeling applications, but require a lot of annotated data and hand-designed features. LSTM-based models are good for capturing sequential dependencies and tend not to be good for capturing long-range context. Transformer models have a strong understanding of context, but they are more computationally expensive and may still be influenced by dataset-specific tokenization effects [4, 6]. Therefore, it is not enough to trust one particular model to provide strong and generalizable PII detection results.

To address these challenges, this study proposes EnsemblePII, a weighted voting approach for robust PII detection in text. The framework combines five diverse detection components: a rule-based model, a dictionary-based model, a Conditional Random Field sequence tagger, a Bidirectional Long Short-Term Memory sequence model, and a transformer-based token classifier. Each component independently processes the input text and produces token-level predictions. These predictions are then merged using a weighted voting aggregator, where the influence of each model is determined by its validation performance for each PII category. This approach enables the framework to combine the strengths of symbolic, statistical, sequential, and contextual models while reducing the limitations of individual methods.

The main contributions of this study are threefold. First, it develops a class-specific weighted voting strategy for combining heterogeneous PII detection models at the token level. Second, it evaluates the proposed framework against its individual components using a unified BIO tagging scheme across seven canonical PII categories. Third, it investigates the effect of domain variation by testing the models on both a general-purpose PII dataset and a domain-specific financial dataset, highlighting practical challenges such as tokenization noise, entity imbalance, missing entity classes, and domain-specific weight calibration.

The rest of this study is structured as follows. Section II discusses previous research on rule-based, machine learning, deep learning, and hybrid methods for PII detection. The EnsemblePII framework and the voting mechanism are presented in Section III. Section IV presents details on datasets, the preprocessing pipeline, implementation, and evaluation metrics. Experimental results are presented and discussed in Section V. Section VI compares the proposed approach with other hybrid PII detection systems. The conclusion of the study is stated in Section VII. Future research directions are presented in Section VIII.

II. Literature Review

Prior research on PII detection in NLP can be grouped into rule-based, machine-learning, deep-learning, transformer-based, and hybrid approaches. Building on these approaches, several studies have explored machine learning and deep learning techniques to enhance the detection and protection of personally identifiable information.

Study [4] conducted a systematic comparative evaluation of machine learning approaches for automated PII detection in unstructured text, assessing rule-based methods, classical classifiers (including SVM), and deep learning models such as BERT. The research evaluated rule-based methods, classical classification techniques (including SVM), and deep learning models (such as Transformers/BERT). The analysis showed that deep learning models, such as BERT, have high accuracy and recall, surpassing the classical methods. However, the authors observed that these better-performing models require substantial computing power and a lot of data.

Study [9] developed an artificial intelligence (AI) model that uses natural language processing (NLP) approaches to identify sensitive personally identifiable information (PII) in unstructured text. The authors applied a pre-trained bidirectional encoder representation from the transformers (BERT) model, re-trained on a balanced dataset including both sensitive and non-sensitive texts, and used cross-validation to obtain accurate results. The results from experiments showed an accuracy of 99.558% and a 7.5% improvement over the previous model, indicating the success of combining NLP with deep learning models for the automatic detection of sensitive data for privacy protection. The research also suggested extending the experiments to more sensitive areas (such as healthcare and finance) and building scalable and privacy-focused versions of BERT for real-world applications.

Study [10] proposed a hybrid approach to detect and anonymize Personally Identifiable Information (PII) in financial records using rule-based Natural Language Processing (NLP) and machine learning. This research sought to protect PII such as name, credit card number, and email addresses from data breaches and ensure compliance with privacy regulations such as GDPR and CCPA. They created a varied synthetic dataset with the Faker library and performed detailed annotations to train a spaCy-based Named Entity Recognition (NER) model. The model proved to be effective with a precision of 94.7%, a recall of 89.4%, an F1-score of 91.1%, and an accuracy of 89.4% on synthetic data, and similar accuracy on real financial reports (an average of 93%). This solution addressed issues of semantic consistency, dealing with unstructured data, and speed. But there were some drawbacks with respect to ambiguous entities and biases in synthetic data, and improvements for multilingual and privacy-preserving capabilities can be made in future research. Their research provides a scalable, reliable, and compliant approach to safeguard sensitive financial data.

Study [11] proposed a rapid system for Personally Identifiable Information (PII) discovery with automatic consent discovery. The study highlighted the importance of protecting personal data in line with international standards like GDPR and PDPA. In the study, the authors introduced "PII RapidDiscover," which uses smart search algorithms and natural language processing, incorporating an enhanced Presidio library for English and Thai documents. It automatically detects PII and links with consent management processes. Through practical experiments, the system's effectiveness and efficiency were shown to be superior to current systems, revealing a holistic approach to improving data security, governance, and legal consent obligations.

Study [12] carried out a systematic evaluation study on hybrid approaches for multilingual Personally Identifiable Information (PII) detection, in which a new hybrid framework, RECAP, was developed. The research sought to solve the

problem of PII detection in low-resource languages, which are linguistically diverse and have little to no official data annotation. The framework combines deterministic regular expressions for static, rule-governed PII patterns with context-aware large language models for NER cases requiring semantic interpretation. The RECAP approach has a modular design that enables support for more than 300 entity types in a three-stage PII refinement process without the need for retraining the model for each language. The academic experiments indicate a substantial advantage for the RECAP approach, where it is 82% better than fine-tuned NER models and 17% better than Zero-shot LLMs in the Weighted F1-score. These results support the effectiveness of the hybrid approach in offering a Scalable and Adaptable framework for PII detection in 13 languages for privacy-compliant applications across languages.

Study [13] proposed a new model for cross-domain PII (Personally Identifiable Information) detection, de-identification, and re-identification using semantic features and ensemble machine learning. Their EESD-PII model combines class-balancing methods, including SMOTE, SMOTE-BL, and SMOTE-ENN, with ensemble classifiers such as Bagging, Boosting, AdaBoost, Random Forest, and Extra Trees, using Word2Vec CBOW representations for contextual text features. To tackle the problems of traditional dictionary-based approaches, as well as cross-domain generalization, the model is trained using multi-domain data from financial, medical, and identification records. The reported results indicate that EESD-PII outperformed existing methods, achieving an accuracy of 99.77% and maintaining performance on imbalanced and heterogeneous data sources. The work is significant in the field of privacy protection within modern data management systems. Table I summarizes the most relevant prior studies on PII detection by comparing their model types, datasets, and reported evaluation metrics.

TABLE I. SUMMARY OF RELATED WORKS

Ref	Year	Type of model				Dataset		Evaluation criteria		
		R-B	ML	DL	Hybrid	1 Dataset	2 Datasets	Best F1-score	Best Recall	Best Precision
[4]	2025	✓	✓	✓	✓	✓	✕	97.9%	97.2%	98.6%
[9]	2024	✕	✕	✓	✕	✓	✕	N/A	N/A	N/A
[10]	2025	✓	✓	✕	✓	✕	✓	91.1%	89.4%	94.7%
[11]	2023	✓	✓	✓	✓	✓	✕	97.39%	97.18%	97.63%
[12]	2025	✓	✕	✓	✓	✓	✕	0.657	0.605	0.729
[13]	2024	✕	✓	✕	✓	✓	✕	99.71%	99.63%,	99.81%
Our Proposed Model		✓	✓	✓	✓	✕	✓	97.49%	97.03%	97.95%

Many hybrid approaches for PII detection have been developed, but they continue to struggle with detecting different forms of PII, unstructured text, and contextual patterns. The majority of existing approaches combine only two methods, such as using rule-based detection combined with machine learning or NLP processing combined with a deep learning model. Although these methods outperform stand-alone systems, they still struggle to detect different types of PII in various text structures. Most current studies also use small data sets, synthetic data, or domain-specific data, making it hard to verify if their models are robust across different domains. Furthermore, some research only provides a subset of evaluation measures, making it harder to compare precision, recall, F1-score, and overall system performance.

To overcome these challenges, this study proposes a novel hybrid model that integrates rule-based patterns, dictionary matching, machine learning, deep learning, and transformer-based contextual classification models in an ensemble approach. This helps broaden the detection scope, as each model can contribute to the overall detection based on their strengths: rule-based approaches support structured identifiers, dictionary matching supports known entities, sequence models support contextual information, and transformer-based models support semantic information. Through a weighted voting approach, the proposed model seeks to enhance consistency and stability in predictions for different PII categories. The proposed model is also evaluated on two datasets to support its generalizability, making the framework more appropriate for detecting PII in various types of unstructured text.

III. ENSEMBLEPII FRAMEWORK

EnsemblePII is a weighted voting ensemble framework that combines five PII detection models working together on the same input text. Each model focuses on a different aspect of PII recognition, and their predictions are combined through a class-specific Weighted Voting Aggregator.

A. System Overview

This method proposes the development and evaluation of a comprehensive PII detection system that is built to balance performance with efficiency. EnsemblePII is a modular ensemble architecture that integrates five heterogeneous PII detection components under a unified per-class F1-weighted voting aggregation scheme. The framework integrates five complementary detection components: rule-based regex pattern matching, dictionary-based gazetteer matching, a Conditional Random Field (CRF) sequence tagger, a Bidirectional LSTM-CRF (Bi-LSTM-CRF) sequence model, and a DistilBERT transformer fine-tuned for token classification. The core of this method is the individual development and evaluation of five different PII detection models. Every model deals with the task from a different perspective, providing unique strengths and weaknesses of the models with respect to accuracy, contextual understanding, and computational cost.

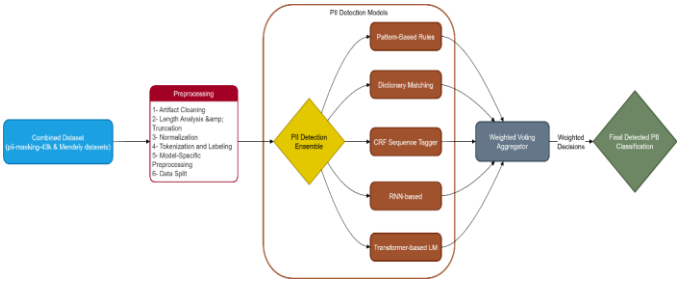


Fig. 1. EnsemblePII system architecture.

Fig. 1 shows the parallel input of the raw text to the five independent detection models (Pattern Rules, Dictionary Matching, CRF, RNN-based, and Transformer-based), each separately.

The results of the five models are combined by the "Weighted Voting Aggregator" that analyzes them with the help of category-specific weights for the final list of classified PII tokens.

B. Component Models

The EnsemblePII framework consists of five complementary detection components, combining rule-based, dictionary-based, statistical, sequential deep learning, and transformer-based approaches for PII detection [4].

1) Rule-based pattern matching (Regex): This component uses regular expressions (regex) to detect the most common PII patterns (such as phone numbers and social security numbers) [14]. The component uses a deterministic rule-based model to target structured PII that appears in consistent, well-defined formats. A large set of regex rules, which are designed to identify structured entities, is matched against the input text. Rules are divided into two main groups: general patterns apply to entities like IP addresses that follow common patterns in their representations; region-specific patterns apply to entities like national identification numbers, which need to be specially created patterns. The technique is able to detect high precision in a well-defined data type.

However, it tends to return false positives, and it does not work well with unstructured or irregular text [4]. Regex-based systems do not consider semantic context, thereby often dealing with poor semantic resolution and issues of surface variation. For example, a pattern that matches only numbers wouldn't correctly "red flag" a number that isn't sensitive.

2) Dictionary matching: This component uses gazetteers and lexical dictionaries of common names, locations, etc., to identify PII [4]. This technique is usually used in conjunction with others in hybrid PII detection systems [10]. Some PII detectors for network traffic, for instance, merge structured pattern matching techniques (e.g., regular expressions) and dictionary/lookup approaches. The technique works well on names of known entities, like first names, last names, and city names, added to the known list. However, its usefulness is highly dependent on the completeness of the underlying dictionaries it uses: it may not recognize entities not included in the dictionaries it uses (e.g., uncommon or misspelled names) and it cannot distinguish between entities that are not included in the dictionary that can be recognized as a person name from those that serve another grammatical role (e.g., common tokens such as Mark may be recognized as a person name in contexts where they do not play that role).

3) Conditional Random Field (CRF) sequence tagger: This component relies on a statistical sequence model that has been trained for PII span recognition based on the presence of contextual features in the surrounding text. It is specifically based on Conditional Random Field (CRF), a structured prediction framework, which is known to be a good baseline in sequence labelling problems [5]. In this setting, data anonymization is formulated as a sequence labeling task, where CRF models assign labels to tokens to identify and categorize personal data entities (such as names, addresses, or dates) and map them to anonymized tags [15]. To improve PII detection, the CRF component models dependencies between adjacent labels within the prediction pipeline. Sequence labeling is one of the tasks where the objective is to convert an input token sequence into an output label sequence of the same length. The CRF model is a stand-alone structured prediction model. It labels tokens by modelling the conditional probability of the whole label sequence conditional on the input token sequence. The CRF represents a classifier that takes into account dependencies between consecutive labels through the use of transition scores, such as B→I and O→I, with the transition B→I being favored over O→I. This is a global optimization over the label sequence to prevent structurally invalid sequences of labels (e.g., I-PER tag immediately after an O tag). This work adds a densely featured CRF model with token lexical features, character n-grams, token shapes, context window features, and gazetteer flags, which allows it to predict with informed decisions without the need for a neural architecture.

4) Bidirectional LSTM (Bi-LSTM) sequence tagger: This component uses a sophisticated recurrent neural network – the Bidirectional LSTM (Bi-LSTM) – to help it get more context. The Bi-LSTM captures both forward and backward information for each sequence by processing them in both directions. This bidirectional result in a more complete contextual representation than a vanilla unidirectional LSTM. The architecture is comprised of cells that can store, update, and reset their internal states as they go through the sequence. The BiLSTM model is merged with a CRF layer to create a hybrid Bi-LSTM-CRF architecture, which is used in various applications [16]. To enhance the Bi-LSTM model, this work adds a CRF layer to it, resulting in the Bi-LSTM-CRF model. In all tables in this study, this model is evaluated and reported as "Bi-LSTM", but it can be understood to be this Bi-LSTM-CRF architecture. The Bi-LSTM module takes in the sequence of tokens, processes them from both directions, and captures more complex context-dependent patterns [16]. Finally, the CRF layer is stacked on top of the Bi-LSTM output, which models the dependencies between adjacent labels and ensures

that only valid label transitions (B→I, I→O, etc.) are produced, giving globally consistent output sequences. It possesses both rich contextual representation and structured output constraint, and thus works well for sequence labeling tasks like PII detection. The hybrid model between two of the above core techniques (recurrent networks and CRF) is shown in Fig. 2.

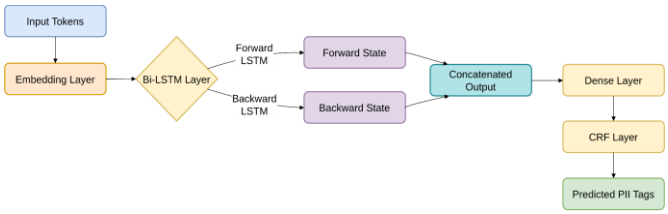


Fig. 2. Bi-LSTM-CRF architecture.

It illustrates the information that passes through the first embedding layer to the two bidirectional LSTMs layers that process the sequence from both directions to get the context, and to the last CRF layer that outputs the tagging decision, together with the learned tag-transition constraints (e.g., an I-NAME tag must follow a B-NAME tag).

5) Transformer-based token classification (DistilBERT): This component uses a transformer-based architecture fine-tuned for token classification to capture contextual representations relevant to PII detection. The design is based on modern NLP models, which are based on the transformer architecture. Transformers are a family of neural network architectures based on self-attention and have achieved strong performance in language representation and language understanding tasks [9]. Unlike models that process tokens strictly in sequence, transformers perform computation in a small, fixed number of steps by applying self-attention to model interactions between all words in a sentence, regardless of their relative positions [8]. The self-attention mechanism enables transformers to find long-distance dependencies, which is important for PII identification tasks in which the meaning of a token is derivable from its relations with other words in the text, rather than just its close neighbors or position.

In this work, the DistilBERT model [17] is fine-tuned for token classification, with a linear classification head to output the final classification of 15 BIO tag classes.

C. Weighted Voting Aggregator

For token Processing, the raw text is processed in parallel by all five models (Pattern Rules, Dictionary Matching, CRF, RNN-based, and Transformer-based). Each model identifies and classifies sensitive tokens, producing its own outputs [13]. Finally, for the Weighted Aggregation, the outputs are merged via the aggregator, which changes the impact of each model based on its evaluation results and category-specific performance. For example, if a phone number token is recognized by both the Regex model and the CRF, the Regex vote may receive greater influence for that class. In the case of a "Name" in an ambiguous context, a Transformer's vote would be given the greatest weight. The aggregator calculates the weighted votes for each token and outputs a final set of classified PII tokens.

It is important to note that per-class F1 weights are derived exclusively from the HF validation set. For entity classes absent from this set, specifically ORG and PHONE, no discriminative calibration signal is available, and all five models are assigned uniform weights for these classes. This constitutes a known methodological limitation: the ensemble weight calibration is domain-specific to the HF corpus, and its cross-domain applicability is not guaranteed.

In this work, the Mendeley validation set is held out exclusively for domain-transfer evaluation; using it during ensemble weight calibration would introduce data leakage into the cross-domain test, and consequently, the HF validation set alone is used as the calibration corpus.

A domain-adaptive weighting strategy, for example, calibrating weights on a held-out mixed-domain validation set or employing meta-learning over validation sets from multiple domains, represents a critical direction for future work.

IV. Materials and Methods

A. Datasets

Two publicly available synthetic datasets were used to train and evaluate the EnsemblePII framework [18, 19].

Dataset A - HF PII43k (ai4privacy/pii-masking-43k): 42,759 rows of text templates filled in various domains. This dataset was filtered, and 42,644 clean rows were kept, with 80% (34,118 samples) used for training, 10% (4,261 samples) for validation, and 10% (4,265 samples) for testing.

Dataset B - Mendeley Synthetic Financial Documents Dataset: 45,000 rows of synthetically generated financial documents rich with PII. From this dataset, 80% (36,004 samples) were used for training, 10% (4,496 samples) for validation, and 10% (4,500 samples) for testing.

TABLE II. Dataset Split Details

Split	Source Dataset	Size	Proportion	Purpose
HF Train	HF PII43k	34,118	80% of HF	Model training (combined)
HF Validation	HF PII43k	4,261	10% of HF	Per-class weight calibration for Ensemble
HF Test	HF PII43k	4,265	10% of HF	Primary held-out evaluation
Mendeley Train	Mendeley	36,004	80% of Mendeley	Model training (combined)
Mendeley Validation	Mendeley	4,496	10% of Mendeley	Internal validation
Mendeley Test	Mendeley	4,500	10% of Mendeley	Domain-transfer evaluation
Combined Training	**Both**	**70,122**	—	**Actual training corpus**

The combined training corpus consists of 70,122 samples (34,118 from HF + 36,004 from Mendeley), which are shuffled and provided together to all trainable models. The evaluation results on two disjoint test sets – HF test set (4265 samples) and Mendeley test set (4500 samples) are reported. Based on this setting, we can evaluate generalization in both domain-specific

and domain-agnostic settings. Table II presents the dataset split configuration used for training, validation, and testing across the HF PII43k and Mendeley financial datasets.

Fig. 3 shows the frequency of the most common PII tag types in the pii-masking-43k dataset. FULLNAME, EMAIL, and NAME are the most common labels, with CITY, URL, and NUMBER following, and the remaining labels, such as USERNAME, PASSWORD, and different types of ADDRESS and ACCOUNT labels, are relatively rare. This means that the data is highly person- and contact-centric, and in fact contains PII.

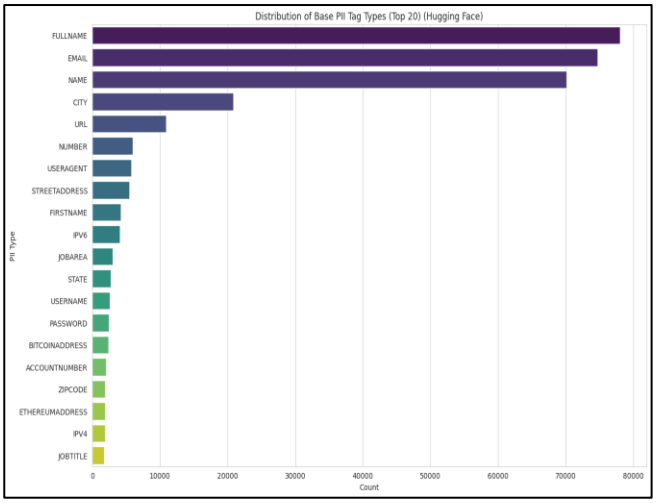


Fig. 3. PII tag distribution in HF dataset.

The frequency of each type of PII tag in the Mendeley data set is presented in Fig. 4. Name, company, and SSN are the most common types, showing that the most frequently occurring types are person or organisation identifiers. Other categories present in the Bar Plot include address, phone, credit card, URL, and email, with a variety of structured and unstructured PII tags. A more balanced distribution across PII types may improve the ensemble model's ability to generalize across sensitive information categories.

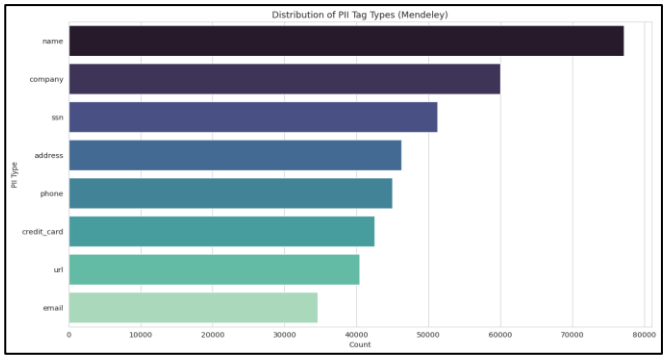


Fig. 4. PII tag distribution in the Mendeley dataset.

B. Unified Label Schema

Because the two source datasets use different entity naming conventions, a unified label schema was constructed to enable joint training. Seven canonical PII entity classes were defined. Table III shows the mapping process used to convert the original dataset-specific entity labels into the unified seven-class PII schema adopted in this study.

TABLE III. LABEL MAPPING SCHEMA

Unified Class	Maps From (HF Dataset)	Maps From (Mendeley)
PER	Name, Firstname, Lastname, Fullname, Displayname	Name, Person
LOC	City, Street address, State, Country, Zipcode, County, Secondary address	Address, Location
ORG	Company, Organization, Department	Company, Org
EMAIL	Email	Email
PHONE	Phone, Phonenumber, Tel	Phone
ID	SSN, Creditcardnumber, Passport, License, IBAN, Bitcoinaddress, Accountnumber, Pin, Vehiclevin, BBAN, Maskednumber, IPV4, IPV6, MAC	Ssn, Creditcard, ID
URL	URL, Domain, Useragent	URL

The entity tags were transformed into BIO format, resulting in a total of 15 labels: O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, B-EMAIL, I-EMAIL, B-PHONE, I-PHONE, B-ID, I-ID, B-URL, I-URL [20].

C. Implementation Setup

For the implementation of the proposed model, a virtual machine environment was used on Google Colaboratory. The accelerators comprise NVIDIA L4 Tensor Core GPUs equipped with 24 GB of GDDR6 memory.

The host CPU back end has about 51 GB of RAM, 107 GB of disk space, and is running Ubuntu Linux (64-bit). For all model training and evaluations, Python 3 was used with Hugging Face Datasets, Scikit-learn, TensorFlow/PyTorch, and the Python re module for regex patterns.

Key libraries used: Hugging Face Datasets [21], Scikit-learn [22], TensorFlow/PyTorch [23], and the Python re module for regex patterns. Table IV summarizes the main hyperparameters and training settings used for each model in the proposed EnsemblePII framework.

TABLE IV. HYPERPARAMETER SUMMARY ACROSS ALL MODELS

Model	Key Hyperparameters	Training Epochs / Iterations
Regex	Deterministic (no training)	N/A
Dictionary	Gazetteer-based (no training)	N/A
CRF	Feature window ±2 tokens; L2 regularization; gazetteer features	Convergence-based (sklearn-crfsuite default)
Bi-LSTM	Hidden dim: 256×2 (bidirectional); Char-CNN embeddings; Word shape features; Adam optimizer; LR = 0.001; batch size = 32	12 epochs
DistilBERT	distilbert-base-uncased; max_length=512; warmup_steps; AdamW; batch size = 32	5 epochs (with early stopping)
Ensemble	Per-class F1 weights from HF Validation set; weighted majority vote per token	N/A (inference only)

To ensure reproducibility, all experiments were conducted with a fixed random seed of 42. Training times on the NVIDIA L4 GPU were approximately 4.2 hours for Bi-LSTM (12 epochs) and 1.8 hours for DistilBERT (5 epochs). Token-level alignment across heterogeneous models was achieved through a unified whitespace tokenisation pre-processing step applied before model-specific encoding, ensuring consistent token boundaries across all five component models.

The Bi-LSTM model includes CNN embeddings for characters and words, and features that focus on words' shape, providing orthographic information that helps identify structured PII like email addresses, identifiers, and phone numbers. The fine-tuned DistilBERT transformer is initialized from the distilbert-base-uncased checkpoint and is then fine-tuned for token classification using a linear classification head with 15 BIO tag classes.

D. Evaluation Metrics

For comparative purposes between various models, their performances are measured in terms of accuracy, precision, recall, and F1-score. These metrics are critical in the characterization of the effectiveness of these models in the detection of PII.

Accuracy: Accuracy represents the overall proportion of correctly classified tokens. However, in PII detection tasks, it can be deceptive, since non-PII tokens greatly outnumber PII tokens.

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (1)$$

where, TP stands for True Positives, TN for True Negatives, FP for False Positives, and FN for False Negatives.

Precision: This is the proportion of the correctly classified tokens predicted as PII, which is the reliability of the positive predictions made by the model.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall: This is the proportion of actual PII tokens predicted by the model. In privacy-preserving applications, a high recall is especially crucial to ensure that most of the sensitive information is detected.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score: This is the harmonic mean of precision and recall, providing a single balanced measure that accounts for both error types.

$$F1\ socre = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

All models are evaluated using the same set of evaluation metrics, calculated over the BIO-tagged labels. The main metric is the weighted F1-score, which addresses class imbalance by weighting each entity class's F1 by its frequency in the test set. Additionally, per-entity precision, recall, and F1 scores are reported for seven standard PII classes: PER (persons), LOC (locations), ORG (organizations), EMAIL, PHONE, ID (identifiers and structured codes), and URL. A prediction is correct only if both the entity boundary and type match the ground truth. This approach is strict, compared to token-level accuracy, and reflects the real-world need to locate PII correctly [24].

V. Experimental Results

A. Test Set Definitions

HF Test Set (4,265 samples): This is a subset from the pii-masking-43k dataset, it includes diverse, general-purpose PII found in conversational and template-based text across multiple fields.

Mendeley Test Set (4,500 samples): This test set is drawn from the Mendeley Synthetic Financial Documents dataset and provides domain-specific financial text, including audit reports, transaction confirmations, and invoices, with PII entities embedded in realistic financial data.

The two test sets enable evaluation of the system's performance across both general-purpose and financial-domain PII detection scenarios.

B. Baseline Model Evaluation

1) Rule-based regex model: The regex model is a pattern-matching baseline model identifying structured PII tokens with the help of a set of tailored regular expressions. No training data is required, and it is simply directly matched to the patterns. Email addresses are detected as top priority; email addresses have very strict syntactic rules followed by URLs, phone numbers, and identifiers. If the model is only based on patterns on the surface, it cannot model context, it cannot resolve ambiguity, not the named entity — a person, place or organization.

The lowest weighted F1 score across all evaluated models is recorded by the Regex model on the HF test set, at 0.1837. A major drawback is that if this model is applied to LOC entities (F1 = 0.000), locations don't have a consistent formatting for pattern-based recognition. The F1 score for the PER class is only 0.2899, which is moderate. This is facilitated by the ability of the regex model to recognize word sequences with capitalized words as candidate names. Similarly, URL has a low F1 score of 0.1048, with many URLs in the HF data set partially masked and thus not matching the regex.

Notably, the EMAIL class, a natural strength for a regex-based model, is also difficult for the model (F1 = 0.0188) on the HF test set. This is due to partially masked email addresses in the dataset (e.g., "##@example.com") that do not fit the standard regex patterns for complete addresses. This observation also accounts for the fact that the model based on the regular expression does much better on the Mendeley test set (weighted F1 = 0.4801) than on the other sets, where email addresses are presented in their full form (EMAIL F1 = 0.9320).

The difference between the results obtained on the HF and Mendeley sets is quite significant: 0.1837 vs 0.4801, which gives us an insight into the fact that the cleaner Mendeley text favors pattern-based models more than the noisier HF text. These results support the decision to include the Regex model as a component of the ensemble rather than as a standalone model [25].

2) Dictionary with regex model: The Dictionary+Regex model adds on entity gazetteers, or a list of known entities from the training data, to the regex baseline. Three gazetteers are built: PER (70,478 person-name tokens), LOC (42,735 location tokens), and ORG (38,111 organization tokens). For structured entities such as EMAIL, PHONE, URL, and ID, the model uses the same regex patterns as in Model 1. Named entities (PER, LOC, ORG) match each input token on the respective gazetteer.

Compared to the pure regex baseline, the Dictionary+Regex model achieves a slight improvement on the HF test set (F1: 0.2635 vs. 0.1837) and its main strength lies in the performance on LOC (F1 = 0.4340 vs. 0.0000). The LOC gazetteer, built from 42,735 training tokens, achieves high precision (0.7566) but moderate recall (0.3042), indicating that many locations in the test sets do not appear in the training data; this is expected in open-world named entity recognition.

These results show the limits of the symbolic methods. Both the Regex and Dictionary models serve as lower-bound baselines, ensuring the need for statistical and deep learning models. Fig. 5 compares the weighted F1-scores of the Regex and Dictionary+Regex baseline models on both test sets.

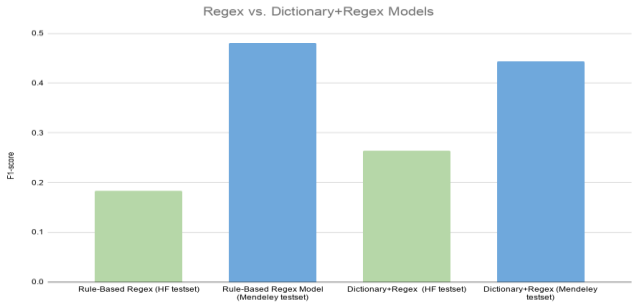


Fig. 5. Weighted F1-score comparison between the rule-based regex and dictionary+regex models on both test sets.

The HF gain (from 0.1837 to 0.2635) contrasts with the drop on Mendeley (from 0.4801 to 0.4432), highlighting the vocabulary-dependency of gazetteer methods.

3) CRF sequence tagger: The Conditional Random Field model shows the transition from symbolic to statistical sequence analysis. It is trained on 70,122 samples, using a different set of features. These features include token-level lexical features, character n-grams, token shapes, context window features, and gazetteer flags for all seven entity classes. The gazetteers for CRF feature includes 7,439 PER, 67,587 LOC, 17,420 ORG, 29,351 EMAIL, 38,156 PHONE, 75,912 ID, and 23,932 URL tokens, all compiled from the combined training data.

The CRF model shows a great improvement over both baselines, achieving an F1 of 0.9469 on the HF Test Set and 0.9639 on the Mendeley Test Set. This represents an approximately 0.68 F1 improvement over the Dictionary model on HF and highlights the effectiveness of statistical feature engineering for PII detection. Five out of seven entity classes achieve F1 scores above 0.84 on the HF Test Set, with EMAIL (0.9868), URL (0.9956), and PER (0.9496) standing out.

The two entity classes ORG and PHONE are completely missing from the HF dataset. As a result, there are no examples of these classes in both the HF Test Set and the HF Validation Set, and all models give an F1 of 0.000 for these classes on HF. This is not a model constraint, but rather a characteristic of the data: with those same models, PHONE is detected with F1 = 0.9463 on the Mendeley Test Set.

However, on the Mendeley Test Set, PHONE achieves F1 = 0.9463, confirming that the issue relates to data quality rather than a model limitation [15].

4) Bi-LSTM sequence tagger: The Bi-LSTM model performs sequence labeling using three embedding channels: word embeddings with a vocabulary size of 265,960, character-level CNN embeddings with 94 unique characters, and word-shape embeddings representing 194 unique shape patterns [26]. These are combined and passed through a two-layer bidirectional LSTM network with a hidden size of 256 per direction, outputting 512-dimensional representations for each token. The model has 35,682,085 trainable parameters, and it is trained for 12 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 32 [27]. Class weights are applied during training, too, to deal with the imbalance between the 'O' (non-PII) class and the rare PII entity classes. Fig. 6 shows the Bi-LSTM training loss over 12 epochs, demonstrating the convergence pattern of the model during training.

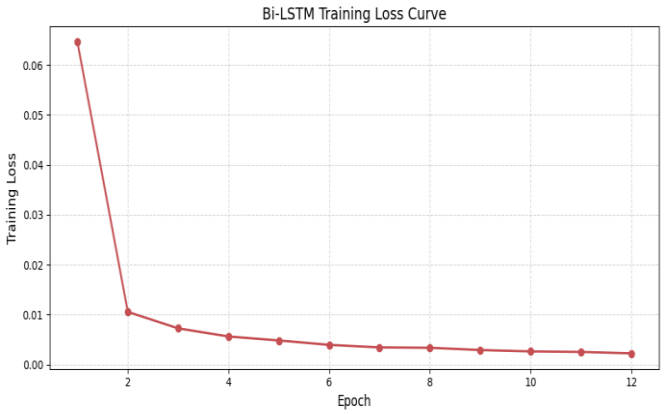


Fig. 6. Bi-LSTM training loss curve over 12 epochs.

Training loss decreases substantially within the first two epochs and continues to decline monotonically thereafter, indicating stable convergence without oscillation or divergence.

The Bi-LSTM model's results reveal an obvious performance discrepancy. On the Mendeley Test Set, it achieves a near-perfect weighted F1 of 0.9959, the highest score of any single model in this study. Four entity classes (ID, PHONE, LOC, URL) achieve near-perfect F1, and even the complex EMAIL class achieves F1 = 0.9991. This strong performance on Mendeley shows the model's ability to learn the consistent patterns of the synthetic financial text, which follows clear template structures.

However, on the HF Test Set, the Bi-LSTM records an F1 of only 0.8433, which is lower than the CRF's 0.9469. This gap is

due to two reasons. First, the HF dataset's tokenization masking again makes it harder to detect EMAIL and URL (F1 = 0.5968 and 0.5590, respectively) because partially masked tokens damage the character patterns the Bi-LSTM needs.

The contrast between 0.8433 (HF) and 0.9959 (Mendeley) shows that the Bi-LSTM is highly sensitive to domain changes. It performs well in well-structured domains but struggles with tokenization noise [28].

5) Baseline summary: Table V compares the weighted F1-scores of the baseline models on the HF and Mendeley test sets, highlighting the relative strengths and weaknesses of each approach.

TABLE V. BASELINE MODEL SUMMARY: F1-SCORE

Model	HF Test F1	Mendeley Test F1	Avg. F1	Key Strength	Key Weakness
Regex	0.1837	0.4801	0.3319	High precision on structured entities (Mendeley)	Cannot detect named entities; breaks on masked tokens
Dictionary+Regex	0.2635	0.4432	0.3534	Adds LOC detection via gazetteer	LOC gazetteer fails on domain shift; ORG/PHONE zero
CRF	0.9469	0.9639	0.9554	Strong cross-domain consistency	ORG/PHONE fails on HF due to masking
Bi-LSTM	0.8433	0.9959	0.9196	Near-perfect on Mendeley	Large performance gap under domain shift

The CRF model stands out as the most robust baseline with an average F1 of 0.9554, achieving strong results across both test sets. The Bi-LSTM reaches the highest single-dataset score on Mendeley (0.9959) but shows weaker performance under domain shift. Both regex and dictionary matching models highlight the limitations of pattern-based approaches, especially in the presence of tokenization noise (in the HF dataset). Their results justify their position as lower-weight contributors in the ensemble system.

C. Transformer Component Evaluation (DistilBERT)

Before evaluating the full ensemble, we discuss the fifth and most powerful individual model, the fine-tuned DistilBERT model. DistilBERT is fine-tuned for token classification using all 70,122 combined training samples [29]. The fine-tuning process was conducted for 5 epochs, with early stopping based on validation F1. The training history is shown in Table VI. Fig. 7 presents the DistilBERT fine-tuning curves, showing changes in training loss and validation F1-score across the five training epochs. Table VI reports the DistilBERT fine-tuning history across five epochs, including training loss, validation loss, and validation F1-score.

TABLE VI. DISTILBERT FINE-TUNING HISTORY 5 EPOCHS

Epoch	Training Loss	Validation Loss	Validation F1
1	0.103743	0.014234	0.9673
2	0.002088	0.010111	0.9757
3	0.001085	0.010821	0.9769
4	0.000555	0.008429	0.9808
5	0.000284	0.007503	0.9804

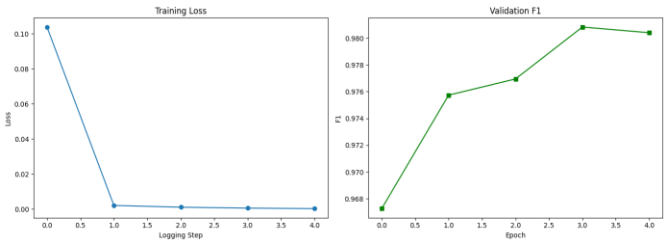


Fig. 7. DistilBERT fine-tuning curves: training loss (left) and validation F1 (right).

The training loss drops sharply within the first epoch. The validation F1 rises from 0.967 to a peak of 0.9808 at epoch 4, showing stable convergence over 5 epochs of fine-tuning.

The DistilBERT Transformer achieves the second-highest F1 on the HF Test Set (0.9744). It scores perfectly on EMAIL and URL, and performs well for PER (0.9781) and ID (0.9632). On the Mendeley Test Set, it reaches a strong F1 of 0.9703, detecting all seven entity classes with no zero scores. Most classes achieve precision close to 1.000, meaning that the model's positive predictions are highly reliable, even though it has lower recall for LOC (0.9224), PER (0.9388), and URL (0.9213).

The main takeaway is that the Transformer's strength lies in its accurate contextual understanding. It almost does not produce false positives, but it may miss some true entities, resulting in lower recall scores. This focus on precision balances the CRF's recall-oriented approach.

D. EnsemblePII System Results

The EnsemblePII system combines predictions from all five models using a dynamic Weighted Voting Aggregator. All five models predict BIO labels for each token in the input text, in the style of voting. Each vote is given a score according to the F1 per entity score of the model on the HF validation set (4261 samples). For example, a transformer has a vote weight of 0.9838, whereas a regex model has a vote weight of 0.3028. For instance, in the case of a PER token, the weight of the vote cast by the transformer is 0.9838, but the weight of the vote cast by the regex model is only 0.3028. In the case of EMAIL, the votes of the transformer and CRF have the greatest impact on the final vote. Table VII presents the per-class validation F1 weights used by the weighted voting aggregator to combine predictions from the five component models.

Table VII displays some of the important properties of the system: For EMAIL and URL, the Transformer and CRF are almost the same in terms of dominant weights. The weight of the Transformer is 0.9838 for PER, followed by CRF (0.9573) and Bi-LSTM (0.9035). The weights for the Regex model and the Dictionary model are less than 0.50 per entity class, implying that both models are auxiliary and not the main decision makers. Note that there is no ORG or PHONE in Table VII. These two

classes are not present in the HF data set, and thus, all models have zero (or near-zero) F1 for the class ORG and PHONE on the HF validation set, and meaningful per-class weights cannot be obtained from the HF validation set for these classes. As a result, for those classes, the ORG and PHONE votes cast by all models are all given the same weight (all models have uniform weight = 1.0), and the fallback scheme is to use majority voting. Table VIII reports the per-entity precision, recall, and F1-score achieved by EnsemblePII on the HF test set.

TABLE VII. PER-CLASS F1 WEIGHT MATRIX ON HUGGING FACE VALIDATION SET

Model	PER	LOC	EMAIL	ID	URL	Overall
Regex	0.3028	0.0000	0.0183	0.0736	0.1272	0.1911
Dictionary	0.2815	0.4577	0.0219	0.0864	0.1379	0.2762
CRF	0.9573	0.9210	0.9811	0.8014	1.0000	0.9501
Bi-LSTM	0.9035	0.8519	0.6325	0.8522	0.6044	0.8491
Transformer	0.9838	0.9635	1.0000	0.9314	1.0000	0.9808

TABLE VIII. ENSEMBLEPII PER-ENTITY RESULTS ON HF TEST SET

Entity	Precision	Recall	F1-Score
EMAIL	1.0000	1.0000	1.0000
ID	0.9357	0.9850	0.9597
LOC	0.9571	0.9470	0.9520
PER	0.9837	0.9703	0.9769
URL	1.0000	1.0000	1.0000
F1-Score	—	—	0.9749

Table IX reports the per-entity precision, recall, and F1-score achieved by EnsemblePII on the Mendeley financial test set.

TABLE IX. ENSEMBLEPII PER-ENTITY RESULTS ON MENDELEY TEST SET

Entity	Precision	Recall	F1-Score
EMAIL	0.9962	0.9925	0.9943
ID	0.9908	1.0000	0.9954
LOC	0.9804	0.9998	0.9900
ORG	0.9368	0.0823	0.1513
PER	0.8214	0.9925	0.8989
PHONE	0.9703	0.9190	0.9439
URL	0.9842	1.0000	0.9920
F1-Score	—	—	0.8433

While EnsemblePII achieves the highest weighted F1 on the HF test set (0.9749), its average F1 across both test sets (0.9091) is lower than that of the DistilBERT Transformer (0.9724), CRF (0.9554), and Bi-LSTM (0.9196). This finding demonstrates that ensemble-based integration improves in-distribution performance but is subject to calibration-induced degradation in cross-domain scenarios, specifically when entity classes present in the target domain (ORG, PHONE) are absent from the calibration corpus.

The ensemble achieves a perfect F1 on EMAIL and URL, matching the Transformer. For the ID class, the ensemble achieved an F1-score of 0.9597. The ensemble's lower ID F1 score (0.9597 vs. the CRF's 0.8014 individual score) reflects the diluting effect of non-zero voting weights assigned to the Regex and Dictionary models for this class (0.0736 and 0.0864, respectively; Table VII), which introduce systematic noise from lower-performing components.

However, on the Mendeley Test Set, the EnsemblePII records a weighted F1 of only 0.8433. This is well below the Bi-LSTM's 0.9959 and the Transformer's 0.9703. The drop is mainly due to the ORG class, which falls to an F1 of 0.1513 on Mendeley. The low ORG F1 on Mendeley (0.1513) is not due to the lack of precision, but rather extremely low recall (0.0823). The ensemble precision for ORG is 0.9368, meaning that approximately 93.68% of predicted ORG tokens are correct. However, since ORG calibration weights are also calculated on the HF validation set, which does not include ORG, all five component models contribute equally (and non-zero) in the ORG predictions. The models that naturally make ORG predictions, via their architecture (probably the Transformer model and the Dictionary model), sometimes predict ORG for ORG tokens, while the remaining models suppress ORG predictions, so the ensemble recall is very low. The number of rare predictions that do pass through is very accurate, which is why the precision is high, and the recall is low, and hence the low F1. ORG and PHONE are completely missing from the HF dataset (validation and test splits), and so no validation weights could be computed for these classes with HF. The ensemble consequently resorts to voting for ORG and PHONE with equal weights when tested on Mendeley.

E. Aggregate Performance Comparison

Table X provides an aggregate comparison of all evaluated models on both test sets, including their rankings and average F1-scores.

TABLE X. MASTER PERFORMANCE COMPARISON: ALL MODELS ON BOTH TEST SETS

Rank (HF)	Model	HF Test F1	Rank (Mendeley)	Mendeley Test F1	Avg. F1
1	EnsemblePII	0.9749	4	0.8433	0.9091
2	Transformer (DistilBERT)	0.9744	2	0.9703	0.9724
3	CRF	0.9469	3	0.9639	0.9554
4	Bi-LSTM	0.8433	1	0.9959	0.9196
5	Dictionary+Regex	0.2635	6	0.4432	0.3534
6	Regex	0.1837	5	0.4801	0.3319

As shown in the aggregate performance matrix, no single model performs best on both test sets. EnsemblePII achieves the highest score on the HF Test Set, with an F1-score of 0.9749, but its performance drops to 0.8433 on the Mendeley Test Set. In contrast, the Bi-LSTM ranks fourth on the HF Test Set with an F1-score of 0.8433, but achieves the best performance on the Mendeley Test Set with an F1-score of 0.9959. The DistilBERT Transformer is the most stable model across both datasets, achieving an average F1-score of 0.9724 and ranking second on both test sets.

When comparing the models by average F1-score across the two test sets, the Transformer is the most consistently reliable model, followed by the CRF, Bi-LSTM, and EnsemblePII. The lower average score of EnsemblePII, 0.9091, compared with 0.9724 for the Transformer, is mainly caused by its weak ORG detection on the Mendeley Test Set. This suggests that the ensemble's reduced cross-domain performance is primarily due to domain-specific weight calibration rather than the ensemble architecture itself. Fig. 8 compares the weighted F1-scores of all evaluated models on the HF and Mendeley test sets.

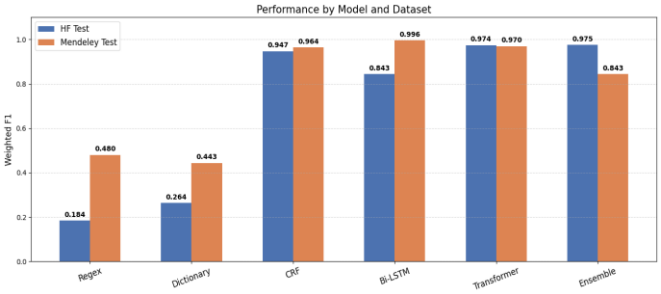


Fig. 8. Weighted F1 bar chart: All models vs both test sets.

Fig. 9 presents the per-entity F1-score heatmap for all models on the HF test set, showing how performance varies across different PII classes.

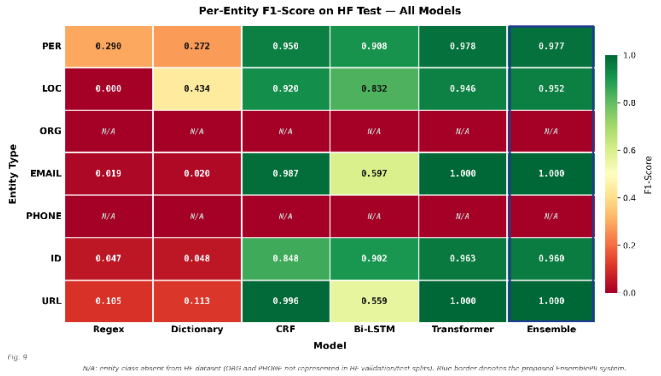


Fig. 9. Per-entity F1 heatmap on HF test set.

Fig. 10 presents the per-entity F1-score heatmap for all models on the Mendeley test set, highlighting model-specific strengths and weaknesses across financial PII classes.

The heatmap places the seven entity classes in rows and the six systems in columns; each cell encodes the F1-score for the corresponding entity-system pair. This visualization offers a quick summary of how each model performed for each class. Transformers and Ensembles show the darkest cells for the EMAIL, URL, PER and ID. However, the ORG and PHONE rows for all models on the HF set are uniformly marked as absent (N/A) in red, reflecting their complete exclusion from the HF dataset rather than a model failure. This CRF column is flatter than the Bi-LSTM column—the CRF shows more uniformity of performance across the entities on the HF data. Contrasting this behavior to the Mendeley heatmap, it is seen that the CRF, Bi-LSTM and Transformer models obtain consistently strong scores. In contrast, EnsemblePII performs well for most classes but shows reduced ORG and PER performance because weaker component models influence the aggregated output for these classes.

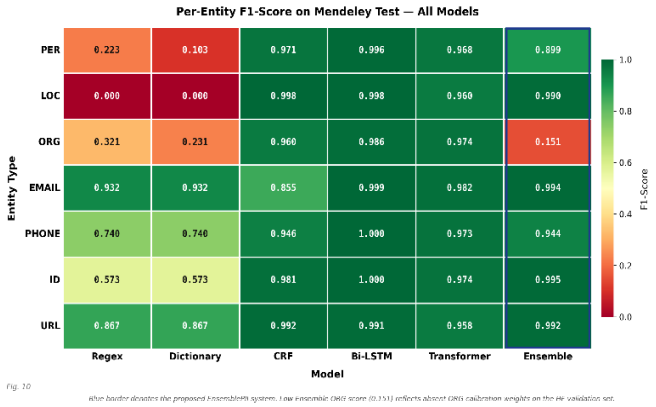


Fig. 10. Per-entity F1 heatmap on the Mendeley test set.

F. Precision–Recall Trade-off Analysis

An important part of evaluating PII detection is the precision-recall trade-off. In privacy protection tasks, recall is usually the more important measure. A missed PII entity represents a false negative and may create a privacy gap, whereas over-prediction produces additional false positives and unnecessary masking. Table XI shows the precision and recall for the three best models on each test set compared to the EnsemblePII framework [30].

TABLE XI. OVERALL PRECISION AND RECALL FOR TOP-3 MODELS

Model	HF Test Precision	HF Test Recall	Mendeley Precision	Mendeley Recall
CRF	0.9543	0.9398	0.9660	0.9624
Bi-LSTM	0.8370	0.8499	0.9948	0.9970
Transformer	0.9776	0.9714	0.9994	0.9424
EnsemblePII	0.9796	0.9703	0.9515	0.8513

The DistilBERT Transformer shows the highest precision on the Mendeley dataset and the highest precision among individual models on HF, while EnsamblePII has the highest HF precision overall, which means that its positive predictions are reliable. On Mendeley, almost every entity that the Transformer identifies is correct (macro precision ≈ 0.9994), even though it misses around 5.8% of true PII entities, resulting in a recall score of 0.9424. The CRF model achieves a balanced precision-recall profile on HF Test, with a difference of only 0.0145 between precision (0.9543) and recall (0.9398). The Bi-LSTM achieves a balanced performance on Mendeley (precision 0.9948, recall 0.9970), showing a good balanced for that domain.

The EnsemblePII shows high precision on HF (0.9796) but a lower score on Mendeley (0.9515). it shows a degraded recall on Mendeley (0.8513), again driven by some baseline collapse for complex entity classes. For applications where recall is critical, the Bi-LSTM on Mendeley-type data or the Transformer on HF-type data would be preferred over the full ensemble.

G. Discussion

The progressive performance improvements from symbolic baselines to the full ensemble system can be attributed to three cumulative factors: 1) feature richness, 2) contextual depth, and 3) ensemble complementarity, each of which is examined in detail below.

Feature richness is most clear in the improvement from the Dictionary model (F1 = 0.2635 on HF) to the CRF (F1 = 0.9469). The CRF's feature engineering, which includes character n-grams, token shapes, and all seven entity gazetteers, gives the statistical model enough information for the final classification signal to reach near state-of-the-art performance without using deep learning. This suggests that, for sequence labeling on well-structured PII data, carefully designed features can remain competitive with neural methods.

Contextual depth explains the Transformer's improvement over the CRF in most individual entity classes. The DistilBERT's self-attention mechanism processes all tokens together, allowing it to determine entity boundaries that depend on sentence-level context. For example, it can tell the difference between a person's name and an organization's name in "John Smith from Smith and Associates." The Bi-LSTM provides intermediate contextual depth. It is stronger than the CRF's fixed context window, but not as expressive globally as self-attention.

The independence of errors by different models (ensemble complementarity) explains the slight improvement of the ensemble over the Transformer on the HF Test. The performance of the CRF in terms of LOC recall (0.9111) is close to that of the Transformer (0.9462), and the ensemble's LOC recall (0.9470) exceeds that of either model.

The effects of the combination of both datasets for training were clear. The CRF model is evaluated on Mendeley with a score of 0.9639, which is far above what is expected from a model trained in this way, given the masking problems with HF data. The Bi-LSTM obtains an outstanding score of 0.9959 on Mendeley, demonstrating its capability of comprehending the structure of synthetic financial text. We hypothesise that DistilBERT's cross-domain stability arises from simultaneous exposure to both HF and Mendeley training distributions, enabling the model to learn domain-invariant contextual representations — a hypothesis that can be tested by training DistilBERT exclusively on HF data and evaluating on Mendeley. The DistilBERT is quite good on both datasets (HF: 0.9744, Mendeley: 0.9703), implying that it learned both distribution styles simultaneously.

An implicit component-contribution analysis is available through the per-class weight matrix in Table VII. The Regex and Dictionary models receive overall validation weights of 0.1911 and 0.2762, respectively, substantially lower than the CRF (0.9501), Bi-LSTM (0.8491), and Transformer (0.9808). These near-zero weights indicate that the two symbolic models exert minimal influence on token-level decisions for most entity classes; the ensemble effectively reduces to a three-model weighted vote among CRF, Bi-LSTM, and Transformer for the majority of predictions. The main contribution of the symbolic components is a precision-oriented signal for highly structured PII patterns such as EMAIL and URL, where deterministic regex and gazetteer matching provide reliable but lower-recall coverage that complements the learning-based models.

However, combined training introduces limitations during the ensemble weight-calibration step. Since ORG and PHONE are completely absent from the HF dataset, there are no validation F1 scores for these classes in the HF validation dataset, and meaningful per-class calibration weights can not be derived for them. This means in this ensemble, there is no calibrated guidance to detect ORG and PHONE on Mendeley, which is a good representation of both classes. This leaves the ensemble unaware of these entity classes, even in the case of evaluating on Mendeley, where they are well defined. This is a problem that could be solved with a domain-aware weighting strategy [31].

VI. Comparison with State-of-the-Art Hybrid Systems

In order to put the performance of EnsemblePII in perspective, the performance of two existing hybrid PII detection systems from the literature is compared to that of EnsemblePII.

Study [9] reported strong results for a BERT-based hybrid model, with 99.558% accuracy and a 99.559% F1-score. Their system, however, is assessed as a binary document-level classification system, in which a classification is performed based on the presence or absence of PII. EnsemblePII tackles a more specific task, the multi-class token-level sequence labelling with BIO tags.

Hence, the results are not directly comparable, but it demonstrates the effectiveness of transformer-based approaches for PII detection.

EnsemblePII attains an F1 of 97.49% on the HF test set, compared to the 91.1% F1 reported by the study [10]; however, direct numerical comparison must be treated with caution because study [10] operates at the entity-span detection level, whereas EnsemblePII addresses the more granular multi-class token-level BIO tagging task, a fundamentally different evaluation granularity. Within this constraint, EnsemblePII also demonstrates a more balanced precision–recall profile (Precision: 97.95%, Recall: 97.03%) than [10] (Precision: 94.7%, Recall: 89.4%). It does not, however, perform as well on a Mendeley financial dataset, where most problems stem from a domain mismatch in ensemble weight calibration, impacting the most important entity classes ORG and PHONE. Table XII compares EnsemblePII with selected hybrid PII detection systems from the literature, while noting differences in architecture, evaluation dataset, and task level.

TABLE XII. COMPARATIVE ANALYSIS: ENSEMBLEPII VS SOTA HYBRID SYSTEMS

System	Architecture Type	Precision	Recall	F1-Score	Evaluation Dataset	Task Level
[9]	NLP preprocessing + fine-tuned BERT	99.564%	99.558%	99.559%	pii-masking-200k (binary)	Document-level binary
[10]	Rule-based NLP + spaCy NER	94.7%	89.4%	91.1%	Synthetic financial (synthetic)	Entity-span level
EnsemblePII (HF Test)	**Weighted voting: Regex + Dict + CRF + Bi-LSTM + DistilBERT**	**97.95%**	**97.03%**	**97.49%**	**HF PII43k (multi-class token)**	**Token-level multi-class**
EnsemblePII (Mendeley)	**Weighted voting ensemble**	**95.15%**	**85.13%**	**84.33%**	**Mendeley financial (multi-class token)**	**Token-level multi-class**

Fig. 11 compares EnsemblePII with selected state-of-the-art hybrid systems using precision, recall, and F1-score values reported in the manuscript.

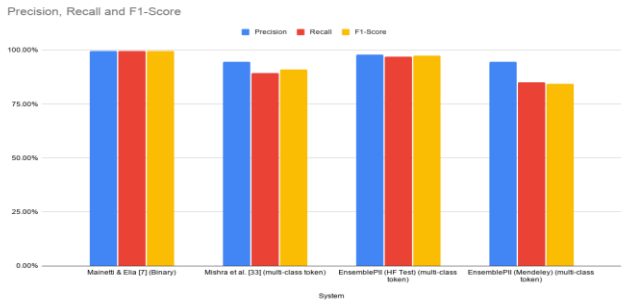


Fig. 11. EnsemblePII vs SOTA: Precision/Recall/F1 Bar.

EnsemblePII outperforms the system in [10] (91.1%) on a more challenging multi-class token-level task on the HF test set with an F1 score of 97.49%. On the Mendeley financial test set, EnsemblePII achieves an F1-score of 84.33%; this lower result is primarily attributable to domain-specific weight-calibration failures for ORG and PHONE entities. Concerning the Mainetti and Elia system: since it is a binary document level task (PII present/not present), it is not appropriate to make a direct numerical comparison. Finally, it should be emphasized that EnsemblePII tackles a much more difficult problem: multi-class token-level sequence labelling with 15 BIO tags, and that it yields strong performance on this more difficult task. First, concerning the complexity of the tasks: EnsemblePII is a multi-class task, token-level classification with 15 BIO tags and 7 entity types, whereas binary detection or single-class entity extraction is simpler. The score of 97.49% F1 on the test set (HF) shows the robustness of the ensemble on general-purpose PII text compared to the 91.1% F1 obtained by the study [10] on a similar entity span task. Second, when it comes to robustness, the two comparison systems do not report on cross-domain generalization. The only system included in the comparison that has been tested on two different types of test sets is EnsemblePII, which used both domain-specific financial documents and general-purpose PII text. This provides more evidence of its real-world applicability. Third, it comes to Precision-Recall Balance, EnsemblePII on HF Test gives better balance (Precision: 97.95%, Recall: 97.03%) than the study [10] (Precision: 94.7%, Recall: 89.4%). This results in EnsemblePII capturing the sensitive entities with fewer false alarms. This twofold benefit is particularly important for applications that have privacy requirements.

VII. CONCLUSION

In this study, we introduced EnsemblePII, a weighted voting ensemble framework for multi-class PII detection in unstructured text. The framework combines five complementary detection approaches: Regex pattern matching, Dictionary+Regex matching, CRF sequence tagging, Bi-LSTM sequence modelling, and DistilBERT-based token classification. The models were evaluated on two test sets representing different domains: the HF PII43k dataset and the Mendeley Synthetic Financial Documents dataset. The results show that symbolic baselines alone are insufficient for reliable PII detection, especially in noisy or masked text. Regex and Dictionary+Regex achieved low weighted F1-scores on the HF test set, while the CRF, Bi-LSTM, and Transformer models provided much stronger performance through contextual and sequence-based learning.

The experimental results indicate that EnsemblePII achieved the best performance on the HF test set, with a weighted F1-score of 0.9749, slightly outperforming the DistilBERT Transformer model, which achieved 0.9744. However, the ensemble did not achieve the best cross-domain performance. On the Mendeley financial test set, its weighted F1-score dropped to 0.8433, mainly because the ORG and PHONE classes were not represented in the HF validation set used for class-specific weight calibration. In contrast, the DistilBERT Transformer was the most stable individual model across both datasets, achieving 0.9744 on HF and 0.9703 on Mendeley, while the Bi-LSTM achieved the highest single-dataset score with 0.9959 on the Mendeley test set. These findings show that ensemble-based PII detection can improve in-distribution performance, but cross-domain effectiveness is contingent on the quality and domain coverage of the calibration data. Across both test sets, the DistilBERT Transformer achieves the highest average F1-score (0.9724) among all evaluated models, making it the most consistently reliable choice when cross-domain generalization is a deployment priority.

VIII. FUTURE WORK

Several directions for future work emerge from this study. First, domain-adaptive weight calibration, for instance, through meta-learning over multiple validation domains or multi-domain validation sets, should be investigated to address the ORG/PHONE calibration failure identified on the Mendeley test set. Second, the framework should be extended to real-world, non-synthetic corpora, such as medical records and legal documents, to assess performance under natural language

variability and annotation noise. Third, multilingual PII detection represents an important extension, as PII formats are language- and culture-specific, and the current framework is English-only. Fourth, the effect of differential weight-thresholding, excluding models below a class-specific F1 threshold from the ensemble vote, warrants systematic ablation. Fifth, privacy-preserving ensemble training, such as federated learning with differential privacy, is an important direction given the sensitive nature of the training data.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Deanship of Graduate Studies and Scientific Research, Taif University, for its support and for funding this work. The authors also thank their colleagues and reviewers for their valuable guidance and constructive feedback throughout this research.

REFERENCES

[1] I. Makhdoom, M. Abolhasan, J. Lipman, N. Shariati, D. Franklin, and M. Piccardi, "Securing personally identifiable information: a survey of SOTA techniques, and a way forward," IEEE Access, vol. 12, pp. 116740-116770, 2024.

[2] V. Garg, "Safeguarding sensitive information: a comprehensive approach to PII anonymization and data masking," International Journal for Multidisciplinary Research, vol. 4, no. 6, art. 21490, 2022.

[3] M. Abreu de Magalhães, "Data protection regulation: A comparative law approach," International Journal of Digital Law, vol. 2, no. 2, pp. 33-53, 2021.

[4] A. Makhambet and A. Moldagulova, "A comparative analysis of machine learning methods for personal information recognition (PII) in unstructured texts," Computing & Engineering, vol. 3, no. 1, pp. 41-52, 2025.

[5] T. Burr and A. Skurikhin, "Conditional random fields for pattern recognition applied to structured data," Algorithms, vol. 8, no. 3, pp. 466-483, 2015.

[6] F. Mortezapour Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models," Journal on Artificial Intelligence, vol. 6, pp. 301-360, 2024.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[8] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," AI, vol. 4, no. 1, pp. 54-110, 2023.

[9] L. Mainetti and A. Elia, "Detecting personally identifiable information through natural language processing: a step forward," Applied System Innovation, vol. 8, no. 2, p. 55, 2025.

[10] K. Mishra, H. Pagare, and K. Sharma, "A hybrid rule-based NLP and machine learning approach for PII detection and anonymization in financial documents," Scientific Reports, vol. 15, art. no. 22729, 2025.

[11] S. Fugkeaw and P. Sanchol, "Enabling efficient personally identifiable information detection with automatic consent discovery," ECTI Transactions on Computer and Information Technology, vol. 17, no. 2, pp. 245-254, 2023.

[12] H. Rajgarhia, S. Gupta, A. Shaik, G. Praveen Kumar, Y. Santhoshraj, S. N. T. Nishitha, and A. Mukherji, "An evaluation study of hybrid methods for multilingual PII detection," arXiv:2510.07551 [cs.CL], Oct. 2025.

[13] P. Kulkarni, C. N. K., and H. R., "An advanced semantic feature-based cross-domain PII detection, de-identification, and re-identification model using ensemble learning," International Journal of Advanced Computer Science and Applications, vol. 15, no. 12, pp. 763-779, 2024.

[14] C. C. Chiu, C. S. Yang, and C. K. Shieh, "A method to improve the accuracy of personal information detection," Journal of Computer and Communications, vol. 11, no. 6, pp. 131-141, 2023.

[15] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, "Survey on RNN and CRF models for de-identification of medical free text," Journal of Big Data, vol. 7, art. no. 73, 2020.

[16] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proc. 2016 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 2016, pp. 260-270.

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108 [cs.CL], Oct. 2019.

[18] ai4Privacy, "pii-masking-43k," Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/ai4privacy/pii-masking-43k

[19] K. Mishra, H. Pagare, R. Bidwe, and S. Mishra, "Synthetic dataset for PII detection and anonymization in financial documents," Mendeley Data, vol. V1, 2024.

[20] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition," in Proc. Seventh Conf. Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142-147.

[21] Q. Lhoest et al., "Datasets: a community library for natural language processing," in Proc. Conf. Empirical Methods in Natural Language Processing: System Demonstrations, Punta Cana, Dominican Republic, 2021, pp. 175-184.

[22] F. Pedregosa et al., "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[23] M. Abadi et al., "TensorFlow: a system for large-scale machine learning," in Proc. 12th USENIX Conf. Operating Systems Design and Implementation, USA, 2016, pp. 265-283.

[24] D. Asimopoulos et al., "Benchmarking advanced text anonymisation methods: a comparative study on novel and traditional approaches," in Proc. 13th International Conf. Modern Circuits and Systems Technologies, Sofia, Bulgaria, 2024.

[25] D. Singh and S. Narayanan, "Unmasking the reality of PII masking models: performance gaps and the call for accountability," arXiv:2504.12308 [cs.CL], 2025.

[26] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in Proc. 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1064-1074.

[27] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in Proc. 3rd International Conf. Learning Representations (ICLR), San Diego, CA, USA, May 2015.

[28] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," MethodsX, vol. 13, 2024.

[29] T. Wolf et al., "Transformers: state-of-the-art natural language processing," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations, Online, 2020, pp. 38-45.

[30] C. Mansfield, A. Paullada, and K. Howell, "Behind the mask: demographic bias in name detection for PII masking," in Proc. 2nd Workshop on Language Technology for Equality, Diversity and Inclusion, Dublin, Ireland, 2022, pp. 76-89.

[31] J. Zhang, "Evaluating machine learning approaches for sensitive data identification: a comparative study of NLP and rule-based methods," Journal of Advanced Computing Systems, vol. 4, no. 7, pp. 26-38, 2024.

Holistic Model of Advanced Analytics to Optimize Organizational Decision-Making

Juan Carlos Morales-Arevalo, Ciro Rodríguez
Faculty of Systems Engineering and Computer Science, Universidad Nacional Mayor de San Marcos, Lima, Perú

***Abstract*—In increasingly data-intensive organizational environments, decision-making processes require analytical frameworks capable of integrating data governance, advanced analytics, and strategic interpretation under a unified structure. This study proposes a holistic advanced analytics model designed to optimize organizational decision-making through the integration of data quality, data integration, analytical capabilities, and data-driven storytelling within a continuous decision-support lifecycle. The proposed model was developed using the Design Science Research (DSR) methodology and structured according to the intelligence, design, and choice phases of the classical decision-making process. The framework incorporates internationally recognized standards and methodologies, including ISO/IEC 25012, ISO/IEC 11179, CRISP-DM, DataOps, and analytics value chain principles, enabling methodological interoperability and adaptive analytical governance. The resulting artifact was conceptually validated through expert judgment involving seven specialists in analytics, business intelligence, and organizational decision-making. The evaluation produced average scores ranging from 3.29 to 4.57 on a five-point Likert scale, with agreement levels reaching 85.71% in the highest-rated dimension. The results indicate favorable perceptions regarding the model's consistency, interpretability, usefulness, and organizational applicability. The proposed model contributes an integrative and adaptive framework that bridges fragmented analytical practices and supports more informed, scalable, and context-aware organizational decisions.**

***Keywords*—Advanced analytics; decision-making; holistic model; Design Science Research; expert judgment**

I. Introduction

In an organizational environment characterized by complexity, uncertainty, and technological dynamism, data-driven decision-making has become a fundamental strategic pillar. However, recent research warns against the illusion that using data alone guarantees effective decisions, highlighting the need for more integrative and adaptive frameworks [1]. Given this reality, there is an opportunity to develop a holistic model of advanced analytics that considers not only data, but also by the four interrelated dimensions influencing decision quality.

Holistic thinking, characterized by the integration of multiple perspectives, has been shown to improve understanding and problem-solving in complex environments, especially when combined with traditional analytical approaches [2]. This approach has been validated in various contexts, from strategic leadership to inventory optimization using artificial intelligence [3]. Empirical evidence also supports the claim that organizations that adopt advanced analytics systems with a holistic orientation achieve significant improvements in organizational performance, innovation, and decision-making [4, 5].

Despite these advances, there remain significant gaps. Many organizations continue to evaluate decisions based on limited and linear criteria without adequately incorporating human-machine collaborative factors [6]. However, integrative models that articulate cultural, cognitive, technological, and social components under adaptive and multidimensional frameworks are lacking [7]. Recent studies emphasize that thinking style, digital orientation, organizational culture, and data connectivity are interdependent variables that, when treated in isolation, limit the impact of analytics on decision-making [8].

Therefore, this study aims to design and validate a holistic advanced analytics model capable of integrating data governance, analytical capabilities, visualization mechanisms, and organizational decision-support processes within a unified adaptive framework. This model is based on robust conceptual frameworks, such as systemic thinking, PDCA (plan, do, check, and act) cycle theory, bi-level optimization under uncertainty, and dual cognitive decision theories [9–11]. At the same time, it is recognized that the adoption of these models must consider the perception of value, resistance to change, and the challenges of practical implementation in different sectors [12].

In summary, this research aims to provide an integrative framework that overcomes the current limitations of fragmented analytics and strengthens organizations' ability to make smarter, more adaptive, and more sustainable decisions.

II. Related Work

A. Data Analysis and Decision-Making Models

Organizational decision-making models have evolved to address complexity and uncertainty in critical contexts. In [13], the authors proposed a model based on Fractional Continuous Fuzzy Sets (FCFS) that allows for a more flexible and sensitive representation of uncertainty using continuous functions instead of discrete values. This approach improves the accuracy of the evaluation of alternatives, such as the selection of industrial robots, where multiple technical criteria are considered. On the other hand, [14] analyzed the decision-making process in healthcare organizations during crises and identified three models: the CEO model (autocratic decision-making), the CEO-advisor model (participation of advisors), and the Top Management Team (TMT) model, which consists of a collaborative management team. These models reflect how leaders adapt their approach according to pressure and uncertainty and prioritize centralization or collaboration. In a similar context, [15] explored the emerging complexity in the

decision-making process of chess players, demonstrating how the accuracy of decisions varies according to the competitiveness and critical nature of positions in the game. This study highlights the relevance of adaptive strategies in optimizing performance in competitive environments, such as minimizing decisiveness metrics. All three studies emphasize the importance of adapting decision models to the organizational context and the nature of the problem, whether technical or strategic.

The Analytics Value Chain model offers a maturity perspective on how organizations transform data into decisions, [16] classify analytical capabilities into four main approaches: descriptive, responds to what happened; predictive, anticipates what is likely to happen; prescriptive, suggests actions based on predictions; and autonomous, automates decisions using artificial intelligence, as shown in Fig. 1. The study [17] is a framework that aligns business analytics capabilities—descriptive, predictive, prescriptive, and autonomous—with Porter's value chain activities, illustrating how these capabilities enhance value creation at different stages of organizational processes. Their study highlighted the strategic application of analytics in primary (e.g., logistics and marketing) and supporting activities (e.g., HR and technology development), emphasizing measurable outcomes in value chain optimization. Building on this, [18] empirically validated the framework using case studies and secondary data, showing how firms leverage analytics for operational efficiency and competitive advantage, particularly in areas such as supply chain management and customer relationship analytics.

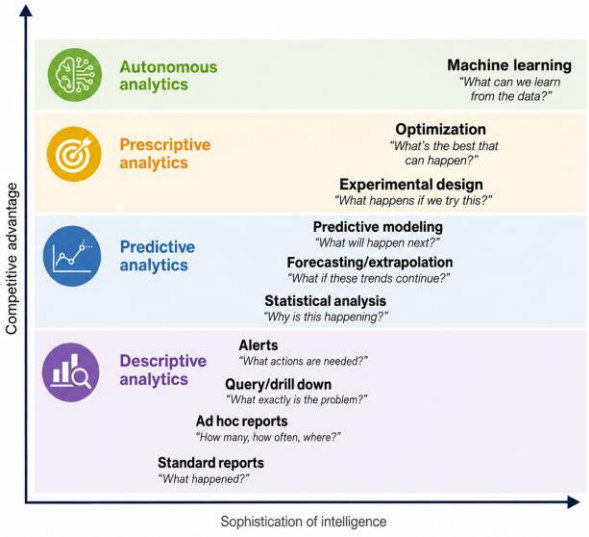


Fig. 1. Potential competitive advantage increases with more sophisticated analytics.

The Data-Driven Decision-Making (DDDM) paradigm emphasizes how the systematic use of data guides each phase of the decision-making process. In organizations that integrate large volumes of data through advanced analytics, DDDM improves the quality of decisions by relying on quantitative evidence rather than intuition or previous experience [19]. This has been demonstrated in areas such as marketing and strategic planning, where data enable the simulation of future scenarios and more informed decision-making [20].

B. Data Design, Analysis, and Management Methodologies

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a widely adopted methodology for structuring data mining projects, offering a systematic approach for extracting actionable insights from complex datasets. In [21], the authors applied CRISP-DM to predict student performance in e-learning environments, emphasizing its six-phase framework—business understanding, data understanding, data preparation, modeling, evaluation, and deployment—to identify at-risk students and reduce dropout rates. Their study highlighted the effectiveness of the random forest algorithm within this framework, achieving an 80% accuracy in predicting academic success. Similarly, [22] utilized CRISP-DM to automate complaint classification in e-commerce, demonstrating its flexibility in handling textual data through natural language processing (NLP) and machine learning. Their model achieved 85% accuracy, underscoring the utility of CRISP-DM for transforming raw data into structured solutions. Both studies illustrate the adaptability of CRISP-DM across domains, from education to customer service, while maintaining rigor in the data analysis and model evaluation (see Fig. 2).

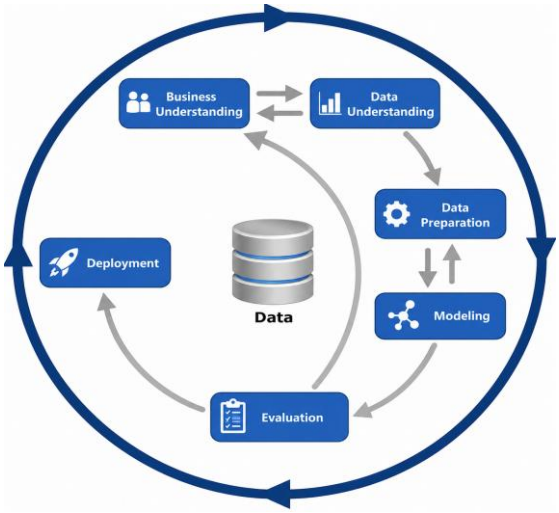


Fig. 2. CRISP-DM process diagram

Regarding the design of business intelligence solutions, Kimball and Inmon's approaches remain relevant as data warehouse modeling frameworks. Kimball proposed a bottom-up approach with an emphasis on dimensional modeling and ease of data access, which has been successfully applied in transactional flow solutions, such as the Data Mart designed for Peruvian bodegas, to analyze POS transactions and improve financial inclusion [23]. Similarly, [24] compares the approaches of Inmon, Kimball, and Data Vault, revealing that each method has an optimal scope of application depending on the project requirements. Inmon's approach is ideal when analytical requirements are not fully defined or when the stability of data sources is a priority, because its centralized 3NF structure ensures long-term consistency. On the other hand, Kimball's model excels in environments where query performance and end-user understanding are critical, especially in data-mart implementations with well-established requirements. In contrast, the Data Vault approach emerges as the most flexible solution for scenarios with multiple dynamic

data sources or agile environments owing to its ability to scale without reengineering and maintain historical traceability. No approach is universally superior, and the choice should be based on factors such as data volatility, project timelines, and end-user needs. Kimball's methodology, shown in Fig. 3, continues to be widely used in various sectors of society. This approach, documented in his data warehouse lifecycle, covers all stages, from business planning to the delivery of analytical value [25].

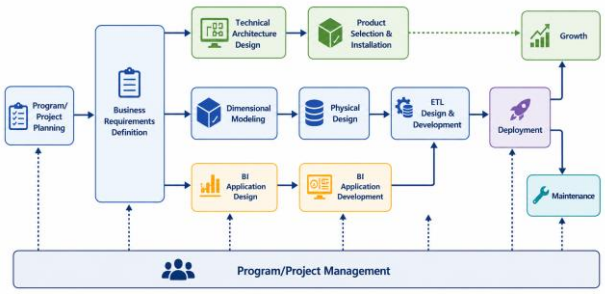


Fig. 3. Business dimensional life cycle diagram.

The Data Management Body of Knowledge (DAMA-DMBOK) framework has gained prominence in data governance due to its comprehensive approach to data quality, metadata management, and integration. A study by [26] proposed a new governance architecture called Horti-IoT, based on DAMA-DMBOK, applied to Dutch horticulture. The results showed significant improvements in the management and integration of data from sensors, images, and manual measurements, highlighting a reduction in analytical effort and an increase in the efficiency of agricultural monitoring.

The DataOps approach, which focuses on automating the data lifecycle and is inspired by Agile, DevOps, and lean manufacturing, has proven highly relevant in environments that require agility and continuous integration [27]. The authors demonstrated that applying this methodology in the feature selection and validation phases can reduce the complexity of datasets without compromising accuracy, achieving 82.32% accuracy with only four key variables. For their part, [28] developed a cloud-independent DataOps framework based on Kubernetes and Apache Iceberg, achieving a 12% reduction in query times through caching and improved scalability in Big Data applications. Their results demonstrated the impact of DataOps on data traceability, information quality, and decision-making efficiency. Together, the studies [27, 28] highlight the potential of DataOps to streamline workflows, shorten development cycles, and ensure robust data governance, thus establishing itself as a key component in modern data-driven organizations.

In complex organizational environments, the DSR methodology provides a rigorous methodological basis for the design and validation of scientific artifacts. This methodology not only emphasizes the creation of innovative solutions but also their systematic evaluation in real contexts. In [30], the authors stated that the DSR approach provides a systematic structure for the design, construction, and evaluation of improvement frameworks in construction processes, thereby ensuring scientific validity and practical relevance. In this study, the authors develop a framework that integrates Lean Construction, Building Information Modeling, and emerging technologies by using DSR to iterate between the design and empirical validation phases.

These methodologies and frameworks demonstrate how technical principles, combined with design and data governance strategies, strengthen organizational analytics models and enable the transition from storage to action with solid, controlled foundations.

C. International Standards for Quality, Security, and Metadata

The implementation of international standards in advanced analytics systems is essential for ensuring data quality, integrity, and security. The main standards related to this research are ISO/IEC 25012, ISO/IEC 11179, ISO/IEC 27001, and ISO 8000, where ISO stands for International Organization for Standardization and IEC for International Electrotechnical Commission, according to recent scientific literature.

The ISO/IEC 25012 standard provides a comprehensive framework for evaluating data quality based on a set of inherent characteristics and system-dependent completeness [31]. In the context of master data management, these characteristics are especially critical, as they ensure the reliability of core business entities that serve as authoritative references across organizational processes. The standard's multidimensional approach aligns with findings in specialized domains; for instance, in healthcare systems, data quality evaluation must consider not only syntactic accuracy, but also semantic precision and temporal relevance to support clinical decision-making. Similarly, industrial applications demonstrate that the ISO/IEC 25012 model effectively addresses measurement challenges through its structured property-based evaluation methodology, which enables organizations to systematically identify and prioritize data quality improvements [32].

With regard to metadata, ISO/IEC 11179 is recognized as the reference framework for metadata governance, particularly in the medical domain. In [33], the authors demonstrated its application in a pragmatic metadata repository that aggregated 466,569 unique medical metadata definitions, facilitating interoperability through standardized data elements. Similarly, [34] emphasized that repositories adhering to this standard enhance semantic interoperability by linking data elements to controlled vocabularies, which is critical for cross-institutional data integration. These studies underscore the role of the framework in ensuring semantic consistency and reliable information exchange in distributed environments.

Regarding security, the ISO/IEC 27001 standard has been widely adopted to protect information assets. According to [35], the framework derived from ISO/IEC 27001:2013 helps organizations identify, assess, and control risks while ensuring compliance with information security laws. Additionally, [36] emphasized its role in balancing privacy and utility in big data analytics, especially in the healthcare sector, strong data protection is essential.

Finally, the ISO 8000 standard plays a critical role in ensuring master data quality, particularly in sectors such as manufacturing and asset management, where consistency and traceability are crucial. As highlighted by [32], ISO 8000

provides a framework for evaluating and certifying master data repositories, emphasizing dimensions such as accuracy, completeness, and compliance—key factors in maintaining reliable "golden records" (p. 795). The study [37] further, ISO 8000's alignment with other standards (e.g., ISO 25012) enables organizations to address both inherent and system-dependent data quality challenges, reducing redundancies and improving interoperability (p. 4). By adhering to these principles, businesses can mitigate operational risks and enhance their decision-making processes. These regulations not only establish best practices but also provide an essential foundation for the design of advanced analytics models focused on data trust, interoperability, and sustainability.

D. Data Visualization for Decision-Making

Effective data visualization allows relevant information to be consolidated in an understandable manner, thus facilitating tactical and strategic decision-making in different organizational and social contexts.

Dashboards and Key Performance Indicators (KPI) have become essential tools for visualizing critical information in real time and facilitating informed decision-making. According to [38], the integration of dashboards into automated processes enabled the visualization of key indicators, such as response time and service level, which increased operational efficiency by 40% through Robotic Process Automation. For their part, [39] highlighted that the implementation of 22 standardized KPIs in dashboards allowed managers to comparatively analyze the performance of different organizational units and optimize strategic decisions. In both cases, real-time visualization not only improved efficiency and accuracy but also strengthened operational control and proactivity in organizational management.

Visual storytelling has established itself as a powerful tool for conveying complex information in an accessible and persuasive manner. In [40], the authors demonstrated the effectiveness of narrative techniques in mental health interventions, where they significantly improved life satisfaction among participants by facilitating emotional engagement and perspective shifts. Similarly, [41] highlighted the role of visual techniques, such as empty shots, in cinematic story-driven methods, showing how their distribution varies across genres to evoke specific emotional and dramatic effects. These studies underscore that effective storytelling combines meaningful visualization with a coherent narrative structure, enhancing understanding and engagement across diverse contexts from healthcare to film. By tailoring visual stories to audience needs—whether through therapeutic group discussions or genre-specific cinematic techniques—storytelling bridges gaps between data, emotion, and action, fostering deeper connections and informed decision-making.

Several studies in the field of Visual Analytics have demonstrated the potential to improve understanding and decision-making in various contexts. For example, [42] developed the Vulnerable Neighborhood Explorer (VNE) tool, which uses interactive visualization and clustering techniques to identify neighborhoods vulnerable to disasters, shows effective results in the analysis of socioeconomic disparities. On the other hand, [43] explored the use of generative AI agents to improve data visualization literacy and found that proactive agents with scaffolding techniques significantly increased student understanding. These studies collectively reveal that although numerous methodologies and frameworks—such as CRISP-DM, DataOps, Kimball, Inmon, and ISO standards—have advanced data management and analytical practices, most remain specialized for particular analytical stages or contexts.

Building upon this body of knowledge, the present study proposes a holistic advanced analytics model that unifies these methodologies under a single adaptive structure. Unlike traditional approaches that require changing frameworks depending on whether the analysis is descriptive, predictive, or prescriptive, the proposed model integrates them coherently, ensuring methodological continuity and adaptability across diverse organizational scenarios. This connection between the reviewed literature and the proposed model establishes a comprehensive foundation for data-driven decision-making in complex environments.

III. METHODOLOGY

This research is based on a combined methodological framework that integrates two complementary approaches: the Decision-Making Process Model and Design Science Research approach. This combination allows us to address the complexity of designing a holistic advanced analytics model to optimize organizational decision-making from both conceptual and empirical perspectives. Within this integrated design, the decision-making model supports the qualitative reasoning phase, while the DSR methodology incorporates quantitative evaluation metrics to ensure empirical validation.

A. Decision-Making Process Model

The classic decision-making process model [44] proposes three fundamental phases: intelligence, design, and choice. As illustrated in Fig. 4, this sequential, but iterative approach provides a structural framework for analyzing complex organizational decisions. Furthermore, [45] empirically examined the distribution of effort across the decision-making phases and found that managers devoted substantially more time to the design phase than to the intelligence and choice phases. This pattern remained consistent across different tasks and computing devices, providing empirical support for the applicability of Simon's model in describing how managers structure problem-solving activities in organizational contexts.

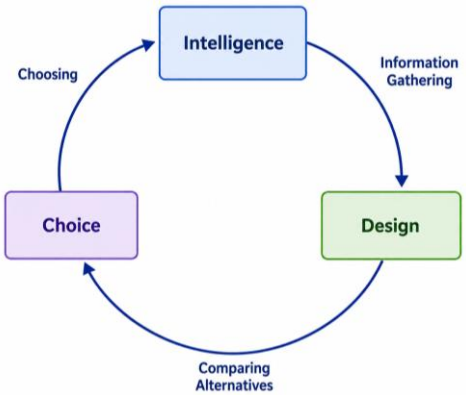


Fig. 4. Three-phase decision-making process model.

1) Intelligence defines this initial phase as the process of scrutinizing the environment to identify situations that require decisions. The author uses the term intelligence in its military sense, emphasizing the systematic collection of critical information that serves as the basis for action (p. 2).

2) Design explains that possible courses of action are generated and analyzed. Design involves creativity (invention of alternatives) and rigor (evaluation of feasibility), constituting a bridge between problem identification and the final selection (p. 2).

3) Choice described this phase as the selection of a specific alternative from among those available. He emphasized that the choice is not merely intuitive but must be based on the criteria established during the previous phases of intelligence and design (p. 2) [44].

B. Approach to Design Science Research (DSR)

The DSR approach provides an epistemological framework that guides the development of a model as a scientific artifact. According to [29] and [46], DSR follows an iterative structure that organizes the development of scientific artifacts through problem definition, objective formulation, artifact design, demonstration, evaluation, and communication, as illustrated in Fig. 5.

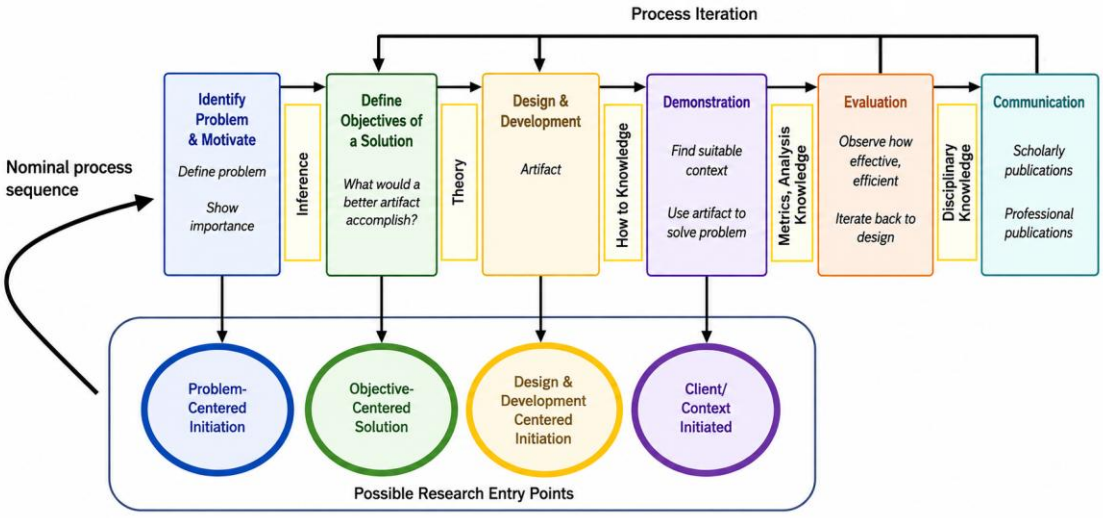


Fig. 5. DSRM process model

1) Identification of the problem and motivation [46] emphasizes that this phase requires clearly determining the research problem and justifying the worth of its solution. Conceptual decomposition of the problem allows its complexity to be grasped, while justification motivates both the researcher and audience. The author highlighted the need to understand the current state of the problem and the relevance of its solution (pp. 28-29).

2) Definition of objectives for the solution, according to [46], the objectives must be rationally derived from the identified problem, considering what is technically feasible. These can be quantitative (e.g., metric improvement) or qualitative (e.g., addressing unresolved problems). The author emphasized the importance of aligning these objectives with the requirements of the artifact, even when the solutions are partial or incremental (p. 29).

3) Design and development [46] describe this phase as the creation of artifacts—constructs, models, methods, or instantiations—that incorporate research contributions. He emphasized that the process requires defining both the functionality and architecture of the artifact, relying on existing theories to substantiate the solution (p. 29). This phase also involves validating the artifact through experimentation, simulations, case studies, or concrete tests which requires a thorough understanding of how to apply the artifact to solve the defined problem. (p. 30).

4) Evaluation [46] proposed comparing the observed results with the objective set using quantitative (e.g., performance) or qualitative metrics(e.g., user feedback). This phase may lead to iterations to refine the artifact or, if successful, to communicate the results (p. 30).

5) According to [46], this final stage consists of disseminating the problem, artifact, and its effectiveness to academic and professional audiences. Publications can be structured following the flow of the DSRM, such as how empirical studies follow a standard format (p. 30).

C. Evaluation Through Expert Judgment

The implementation of Expert Judgment as an evaluation phase in advanced analytics models is justified by its ability to validate the applicability and robustness of the proposed solutions in complex organizational contexts. As demonstrated by the study [47], in the validation of the ACDGE model for personalized learning pathways, Expert Judgment enables for the evaluation of content validity and internal consistency of a model through the participation of specialists with multidisciplinary experience (e.g., academics and professionals with over 30 years in the field). This approach not only ensures the relevance of the model's components (such as the

implementation phase, where significant agreement was obtained with p=0.007) but also identifies areas for improvement, such as simplifying deliverables or incorporating teacher-student co-design.

However, [48] highlighted that expert judgment is key in scenarios with high technical uncertainty, such as the selection of technologies for the reuse of mining waste. Although their study revealed limitations in the distinguishability of alternatives when using questionnaires alone (distinguishability index <0.5), it highlighted that structured interviews with mineralogical design experts improved data quality by clarifying discrepancies and providing technical insights (e.g., chemical process optimization or market viability). This aligns with the need to evaluate advanced analytics models in organizational environments, where data heterogeneity and a lack of historical information require expert validation to reduce bias and ensure practical solutions.

To validate the proposed model, a questionnaire was developed based on instruments recognized in the scientific literature, which allows for the evaluation of key dimensions, such as usefulness, ease of use, understanding, and organizational applicability:

- Technology Acceptance Model (TAM): In [49], the authors state that TAM is widely used to predict the acceptance of technologies, with its core constructs being "perceived usefulness" and "ease of use", which are fundamental for understanding the intention to adopt technology.
- System Usability Scale (SUS): According to [50], SUS allows for quick and reliable measurement of a system's usability, considering elements such as comprehension and ease of interaction.
- Delphi method and expert judgment: In [51], the authors highlighted that the Delphi method allows for the structuring of a systematic process to reach consensus among experts and is highly effective in the evaluation of conceptual models and organizational decisions.

For the content validation of the holistic advanced analytics model, a panel of seven (7) experts was selected. This panel size is consistent with the recommendations of Lynn (1986) [52] and Escobar-Pérez and Cuervo-Martínez (2008) [53], who suggest that groups ranging from five to ten experts are methodologically adequate for content validation studies, ensuring sufficient representativeness and rigor in the assessment of complex constructs while avoiding unnecessary redundancy. The experts were selected through purposive sampling based on their academic qualifications, professional experience in analytics, business intelligence, organizational decision-making, and digital transformation. Each expert conducted the evaluation independently using a structured questionnaire, reducing the potential influence of group bias and strengthening the reliability of the validation process.

IV. Results

This section describes the phases proposed by the DSR methodology to rigorously structure the process of designing, constructing, and validating the proposed model. In particular, the artifact development is detailed through the six defined phases, encompassing everything from problem identification to model communication. Given that one of the key phases of DSR is evaluation, this methodological moment will be used to apply expert judgment as a validation mechanism by collecting specialists' perceptions of the quality and usefulness of the model. Finally, the alignment of each of the DSR phases with the stages of the decision-making process is presented, which reinforces the methodological consistency and practical applicability of the model in the real context of advanced analytics.

A. Identification of the Problem and Motivation

Various studies have shown that organizations face significant challenges in achieving effective decision-making due to fragmentation between data analysis models, the absence of integrated quality and security standards, and poor coordination between key performance indicators and analytical visualization strategies. These limitations result in poorly structured decision-making processes, with little technical support and low information traceability.

Based on this, it is evident that there is a need for an integrated solution that combines good data design and governance practices, advanced analytical tools, and visual resources to strengthen the decision-making process. This justifies the formulation of the following research question:

How can organizations optimize decision-making using a holistic model of advanced analytics?

B. Define the Objectives of a Solution

Based on the identified problem and the need to optimize organizational decision-making through a comprehensive approach to advanced analytics, the main objective of this work is to design and validate a holistic model that integrates the different levels of analytics (descriptive, predictive, prescriptive, and autonomous), as well as processes, standards, and data management methodologies, to strengthen DDDM.

This objective responds to various limitations detected in the literature, such as:

- Fragmentation between data analysis and management models,
- The absence of a methodological structure that integrates quality, security, and governance,
- Limited use of advanced visualization for interpreting critical data,
- And a lack of alignment between the analytical results and actual decision-making processes within organizations.

Therefore, the proposed solution addresses these gaps through the construction of a holistic analytical framework that unifies organizational context analysis, data quality governance, semantic integration, advanced analytical capabilities, and data-driven storytelling within a continuous decision-support lifecycle. The artifact was designed following DSR principles and conceptually validated through expert evaluation to ensure methodological consistency and organizational applicability.

C. Design and Development: Artifact

The design of the holistic advanced analytics model resulting from this research is shown in Fig. 6, which details the components, flows, and actions that comprise the developed artifact in a sequential and functional manner.

The Organizational Context, Table I, shows a sequence of key activities aimed at establishing strategic foundations, analytical requirements, and understanding of the organizational environment prior to working with data. These activities ensure that the advanced analytics process is aligned with business objectives, considering actual needs and available data sources.

The Data Cleaning and Quality, Table II, outlines the steps necessary to ensure that the data used is accurate, consistent, and useful. This stage is based on standards such as ISO 25012 and ISO 8000, which guide data quality management and employ techniques such as profiling and cleaning that are essential for reliable analytics.

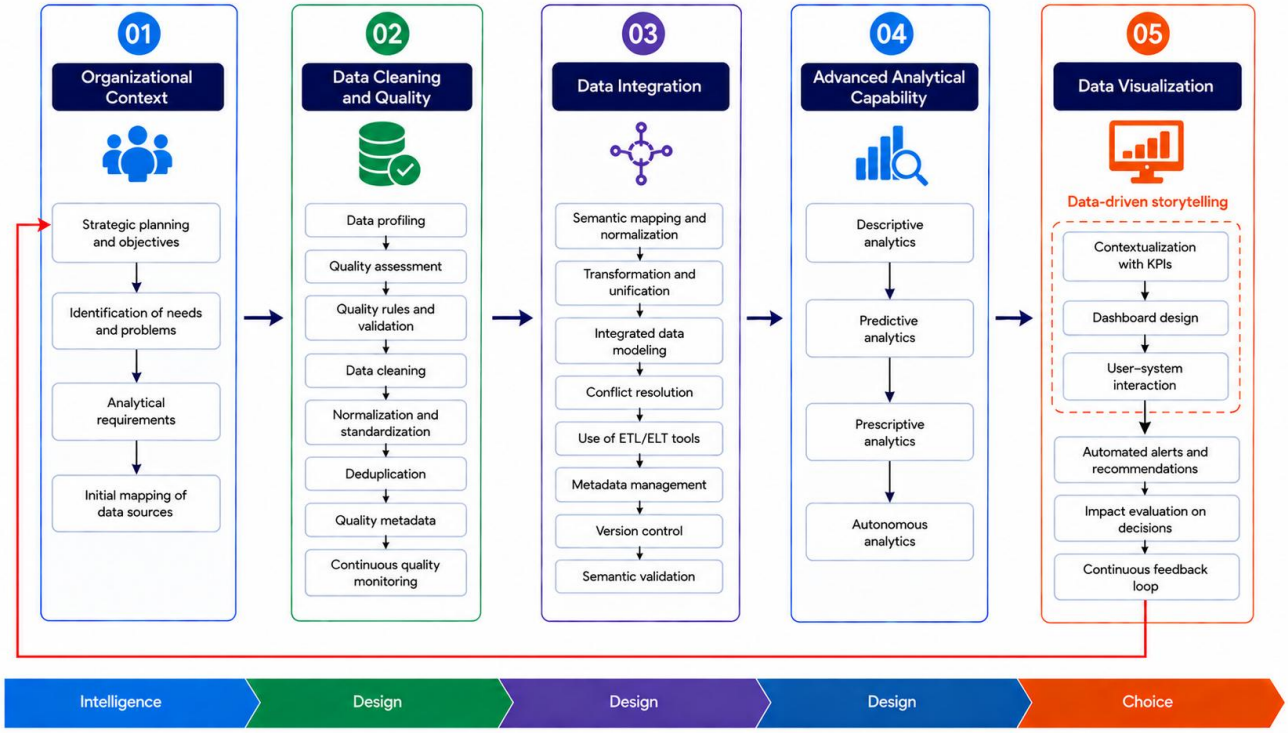


Fig. 6. Holistic advanced analytics model.

TABLE I. ORGANIZATIONAL CONTEXT

Activity	Purpose	Referenced Frameworks / Standards
Strategic planning and objectives	Define the strategic vision and goals for analytical implementation.	Data-Driven Decision-Making (DDDM)
Identification of needs and problems	Detect issues and opportunities for analytical intervention.	Process Model for Decision-Making
Analytical requirements	Establish analytical capabilities required to meet business needs.	DAMA-DMBOK
Initial mapping of data sources	Identify and list internal and external data sources.	DAMA-DMBOK, ISO/IEC 11179

TABLE II. DATA CLEANING AND QUALITY

Activity	Purpose	Referenced Frameworks / Standards
Data Profiling	Analyze data to understand its structure and quality.	DAMA-DMBOK, ISO 8000
Quality Assessment	Evaluate completeness, accuracy, and consistency of data.	ISO/IEC 25012, DAMA-DMBOK
Quality and Validation Rules	Define and apply data validation checks.	ISO/IEC 25012, Kimball
Data Cleaning	Correct errors, remove duplicates, handle missing values.	Crisp-DM, ISO 8000
Normalization and Standardization	Ensure data is in a consistent format.	ISO/IEC 11179, Kimball
Consolidation of Duplicates	Merge duplicate records for consistency.	DAMA-DMBOK
Quality Metadata	Document quality rules and attributes.	ISO/IEC 11179
Continuous Quality Monitoring	Ongoing tracking of data quality KPIs.	DataOps, ISO 8000

TABLE III. DATA INTEGRATION

Activity	Purpose	Referenced Frameworks / Standards
Semantic mapping and normalization	Align and translate terms across sources.	ISO/IEC 11179
Transformation and unification	Standardize formats and structures.	Kimball, Inmon
Integrated data modeling	Create a unified data representation.	Inmon, DAMA-DMBOK
Conflict resolution	Resolve data inconsistencies.	Kimball
Use of ETL tools	Extract, transform, and load data.	Kimball, DAMA-DMBOK
Metadata management	Organize metadata for integration and retrieval.	DAMA-DMBOK, ISO/IEC 11179
Version control	Track and manage changes to data.	DataOps
Semantic validation	Ensure integrated data meets defined semantics.	ISO/IEC 11179

The Data Integration, Table III, shows how the data from multiple sources are consolidated, transformed, and unified. This process is based on practices such as those of Kimball and Inmon for integration and modeling, and is supported by ETL (Extract, Transform, Load) tools and standards such as ISO/IEC 11179 for semantic and metadata management.

The Advanced Analytical Capability, Table IV, details the different advanced analytical capabilities aligned with the Data Analytics Value Chain. These capabilities enable the extraction of value from data using specific algorithms and models, according to the level of analytical maturity.

The Data Visualization, Table V, describes the activities necessary to transform the analytical results into understandable visual elements, aligned with the principles of visual analytics, storytelling, and data-driven decision-making. This phase reinforces organizational action through dashboards, KPIs, and continuous visual feedback.

TABLE IV. ADVANCED ANALYTICAL CAPABILITY

Activity	Purpose	Referenced Frameworks / Standards
Descriptive Analytics	Understand past behavior using statistical summaries.	Value Chain - Business Understanding; Algorithms: Histograms, Clustering
Predictive Analytics	Predict future outcomes using models.	Value Chain - Data Mining; Algorithms: Regression, Decision Trees
Prescriptive Analytics	Recommend actions based on predictions.	Value Chain - Deployment; Algorithms: Optimization, Simulation
Autonomous Analytics	Enable systems to decide and act independently.	Value Chain - Automation; Algorithms: Reinforcement Learning, Deep Learning

TABLE V. DATA VISUALIZATION

Activity	Purpose	Referenced Frameworks / Standards
Data-driven storytelling	Communicate insights through narratives.	Storytelling for Data, Visual Analysis
Contextualization with KPIs	Present insights aligned to performance goals.	DAMA-DMBOK
Dashboard design	Visualize key metrics clearly.	Visual Analytics, Kimball
User–system interaction	Enable users to explore and interact with data.	Visual Analysis, Human-Centered Design
Automatic alerts and recommendations	Provide proactive, rule-based guidance.	Prescriptive Analytics
Impact assessment on decisions	Evaluate analytical influence on outcomes.	Process Model Decision-Making

D. Demonstration

In order for the proposed artifact to demonstrate its applicability, we present an alignment between the phases of the holistic model of advanced analytics and the stages of the classic model of the decision-making process, which include three key moments: intelligence, design, and choice. This correspondence, as shown in Table VI, allows us to observe how the developed model offers structured support to each stage of the decision-making process.

In particular, the initial phase, called the Organizational Context, is designed to adapt to any type of organization, regardless of size, industry, or level of analytical maturity. Its purpose is to clearly establish needs, available data sources, and plans for future actions.

In addition, the model was evaluated by expert judgment, with the participation of seven professionals with experience in data analytics, visualization, and organizational decision-making. The results obtained in this evaluation, which are presented in the next phase, reflect positive acceptance in terms of the model's usefulness, applicability, and methodological alignment.

E. Evaluation (Validation)

A structured questionnaire was administered to a group of experts in advanced analytics and decision-making to evaluate the consistency, applicability, and usefulness of the developed holistic model.

Table VII shows the set of questions asked along with the criteria evaluated for each question.

TABLE VI. ALIGNMENT OF THE PHASES OF THE PROPOSED MODEL AND THE DECISION-MAKING MODEL

Holistic Model Phase	Functional Description	Decision-Making Phase
Organizational Context	Establishes strategic goals, identifies needs, defines analytical requirements, and maps data sources.	Intelligence
Data Cleaning and Quality	Ensures data reliability through profiling, validation, cleaning, standardization, and quality monitoring.	Design
Data Integration	Integrates and harmonizes data using ETL tools, metadata management, and semantic validation.	Design
Advanced Analytical Skills	Applies descriptive, predictive, prescriptive, and autonomous analytics to explore alternatives.	Design
Data Visualization	Presents analytical insights through storytelling, KPIs, dashboards, and user interaction for decision support.	Choice

TABLE VII. QUESTIONS FROM THE VALIDATION INSTRUMENT AND CRITERIA EVALUATED

Nº	Question	Criterion evaluated
1	The proposed model adequately reflects the key phases of a data-driven decision-making process.	Functional relevance
2	The phases of the model are clearly defined and understandable for implementation in an organization.	Structural clarity
3	The model adapts to different organizational contexts without the need for structural changes.	Organizational flexibility
4	The model appropriately integrates recognized frameworks and methodologies (e.g., ISO 25012, DSR, TD Process Model).	Methodological alignment
5	The relationships between the phases of the model are consistent and logically sequential.	Logical consistency
6	The model can be replicated by other researchers or professionals with minimal ambiguity.	Reproducibility
7	The analytical capabilities described (descriptive, predictive, prescriptive, autonomous) are correctly classified.	Functional classification
8	The inclusion of the visualization phase (KPIs, storytelling, dashboards) contributes to the understanding and use of the results.	Interpretative capacity
9	The model adds value to organizational decision-making from a holistic and integrative approach.	Organizational added value
10	I believe that the proposed model is useful and applicable for improving decision-making processes based on advanced analytics.	General assessment of usefulness

Table VIII shows the results of seven experts in the fields of data analysis, visualization, and business intelligence who answered the questionnaire. Each question was rated on a Likert scale from 1 to 5, where 1 represents "strongly disagree", and 5 represents "strongly agree". Three indicators were calculated for each question: average, standard deviation, and percentage of responses equal to or greater than four.

TABLE VIII. RESULTS OF MODEL VALIDATION ACCORDING TO EXPERT JUDGMENT.

Question	Average	Standard deviation	Percentage ≥ 4
Q1	3.71	1.03	57.14%
Q2	4.14	1.12	71.43%
Q3	3.57	0.49	57.14%
Q4	3.71	0.70	57.14%
Q5	3.57	1.18	57.14%
Q6	3.71	0.88	71.43%
Q7	3.71	0.88	71.43%
Q8	4.57	0.73	85.71%
Q9	3.29	1.03	42.86%
Q10	3.71	1.03	57.14%

1) Average: The average score for each question indicated the overall degree of acceptance of each dimension evaluated in the model. Most of the averages were between 3.5 and 4.5, which are positive and shows a favorable trend toward the usefulness and applicability of the proposed holistic model.

2) Standard deviation: This value measures the variability of the responses. Lower values indicate consensus among the experts. For example, question Q3 had a low deviation (0.49), suggesting that experts closely agreed with their responses. Conversely, a deviation of 1.18 (Q5) indicated greater dispersion and possibly divergent views on that dimension.

3) Percentage ≥ 4: This indicator represents the degree of acceptance (high ratings). It can be seen that:

a) Q8: Visualization and interpretation of results had the highest level of acceptance (85.71%), reflecting a high positive consensus.

b) Q9: Obtained the lowest agreement percentage (42.86%), which provided valuable feedback for refining the organizational adaptability dimension of the proposed model. This result contributed to the iterative improvement of the artifact by identifying the need to strengthen contextual flexibility and clarify the integration mechanisms across heterogeneous organizational environments. From a DSR perspective, this feedback was considered a constructive input for the evolution and maturation of the proposed framework.

F. Communication

The Communication phase of the DSR methodology is evidenced through the presentation and dissemination of the results of the developed holistic model, as part of a doctoral research study. The results of the questionnaire administered to seven experts indicate that the proposed model is conceptually coherent, methodologically sound, and relevant to the stated objectives. Most of the indicators per question are above the

recommended acceptance thresholds, with averages higher than 3.5 and agreement levels (ratings of 4 or 5) exceeding 60% in most of the evaluated dimensions.

These findings support the validity and applicability of the model in real contexts, constituting a contribution expected to benefit the academic and professional community interested in data-driven decision-making processes based on advanced analytics. The publication of this scientific study serves as a formal dissemination mechanism, thereby fulfilling the principles of the Communication phase of the DSR methodology.

V. Discussion

A comparison of the proposed holistic model with the reviewed literature clearly shows that existing contributions address only partial dimensions of the analytical process and fail to provide an integrative structure that articulates data quality, methodologies, governance, analytical capabilities, and effective decision support. The holistic character of the proposed model lies in its cyclical and adaptive structure, which integrates the analytical process from descriptive capabilities to autonomous decision-support mechanisms, while maintaining continuous feedback for organizational improvement [16–18], [27–29]. Validation with experts supports this reading, reflecting recognized strengths in visualization, structural clarity, and reproducibility of the model, as well as areas where challenges remain, particularly in organizational adaptability

Firstly, models based on uncertainty representation, such as the FCFS approach [13], provide an important advance for evaluating alternatives in highly technical contexts. However, their application is restricted to the methodological level of calculation, without incorporating data lifecycle management mechanisms, quality standards, or strategic elements that enable their use within complex organizational processes. The proposed model extends these approaches by structuring an adaptive analytical lifecycle that integrates governance standards, semantic interoperability, data quality management, analytical scalability, and decision-oriented visualization mechanisms under a unified framework.

On the other hand, crisis decision-making models proposed in the healthcare context, such as the CEO, CEO–advisor, and TMT approaches [14], emphasize leadership styles and governance in high-pressure environments. While they provide relevant insights into decision-making dynamics, these models do not consider the data engineering processes or analytical support necessary for evidence-based decisions. In contrast, the holistic model articulates the classic decision-making process with data analysis methodologies, governance standards (ISO 25012, ISO/IEC 11179, ISO 8000, ISO/IEC 27001), and visualization mechanisms geared toward decision-making.

Similarly, evidence of emerging complexity in human decisions, such as that observed in the analysis of chess players' behavior [15], demonstrates the importance of adaptability and situational reasoning. Nevertheless, these studies are limited to specific domains and do not constitute frameworks applicable to the organization of an advanced analytics process. The present model integrates the principle of adaptability through continuous feedback, scaled analytical capabilities, and permanent monitoring mechanisms, expanding its application to multiple sectors.

Likewise, frameworks such as the Analytics Value Chain [16] and proposals linking analytical capabilities to the organizational value chain [17, 18] have made it possible to conceptualize the progression from descriptive analysis to autonomous analytics. However, these approaches focus on the functional classification of capabilities, without offering a methodological roadmap that articulates standards, data quality, semantic integration, or evaluation mechanisms. The proposed model extends these contributions by structuring operational phases that include cleaning, integration, modeling, and visualization, along with formal quality assurance mechanisms based on ISO 25012 and ISO 8000.

In terms of methodology, CRISP-DM [21, 22] continues to be a widely used standard, especially in data mining projects. Its main limitation lies in its technical approach and its lack of integration with governance models, international standards, or organizational decision-making processes. For its part, DataOps [27, 28] introduces automation, version control, and continuous orchestration, but focuses on pipeline engineering rather than decision support. The holistic model combines both approaches under a higher framework based on DSR, which allows for the integration of design, validation, communication, and continuous improvement.

Finally, the expert validation results reinforce the conceptual coherence of the model. The high acceptance rate of the visualization phase (85.71% of responses ≥4) confirms the importance of translating analytical results into actionable knowledge. In contrast, the lower agreement levels obtained for functional relevance (Q1, 57.14%) and holistic organizational value (Q9, 42.86%) provided valuable input for refining the proposal. Regarding Q1, expert feedback revealed the need to make the relationship between the analytical process and organizational decision-making more explicit. Consequently, the model was refined by incorporating a transversal representation of decision-making throughout the analytical lifecycle, strengthening the connection between analytical capabilities and organizational outcomes. Similarly, the results obtained for Q9 highlighted the importance of reinforcing the continuous improvement perspective. As a result, a cyclical feedback mechanism was incorporated, connecting the data visualization stage with the organizational context and enabling iterative refinement across the analytical lifecycle. These refinements strengthen the model's holistic character by integrating not only the complete advanced analytics cycle but also continuous organizational learning, adaptation, and decision support.

VI. Conclusion

This study proposed and validated a holistic advanced analytics model aimed at optimizing organizational decision-making through the integration of data governance, semantic integration, analytical capabilities, visualization mechanisms, and data-driven storytelling. Unlike fragmented analytical approaches, the proposed framework structures the analytical process as a continuous and adaptive decision-support lifecycle aligned with the intelligence, design, and choice phases of classical decision-making theory.

The application of DSR enabled the iterative development and conceptual validation of the proposed artifact. Expert evaluations indicated favorable perceptions regarding the model's structural coherence, interpretability, and organizational applicability, particularly in the visualization and analytical integration phases.

The proposed framework provides to the current body of knowledge by providing an integrative structure capable of connecting data quality standards, integration methodologies, advanced analytics, and decision-support visualization within a unified organizational context. Furthermore, the incorporation of continuous feedback mechanisms strengthens the adaptive capacity of the model in dynamic and data-intensive environments.

VII. Future Works

Future research will focus on implementing and testing the proposed holistic advanced analytics model in real organizational environments to evaluate its practical effectiveness, scalability, and adaptability across different sectors. Such empirical validation will provide additional evidence regarding the model's contribution to organizational decision-making and its applicability in diverse operational contexts.

The next stage of this ongoing research involves developing a computational tool that integrates all phases of the proposed model, facilitating its operational deployment, automation, and scalability. Comparative studies with existing analytical and decision-support frameworks will also be conducted to assess the relative advantages, limitations, and organizational impact of the proposed model, as well as to identify best implementation practices.

Additionally, the incorporation of emerging technologies—such as generative artificial intelligence, the Internet of Things (IoT), and blockchain—is proposed to expand the model's analytical, governance, and automation capabilities. Future studies will also explore the integration of cognitive analytics capabilities to enhance contextual reasoning, knowledge discovery, interpretation of unstructured information, and human-centered decision support in complex organizational environments.

Furthermore, the model will be adapted and tested within specific sectors, including healthcare, education, and logistics, to validate its contextual flexibility and sector-specific applicability. Finally, future work will focus on the computational operationalization of the proposed framework through the development of an interoperable analytical platform capable of automating the model lifecycle in real organizational environments and enabling longitudinal evaluations of its impact on analytical maturity, data governance, digital transformation, and organizational performance.

References

[1] Szukits Á., "The illusion of data-driven decision-making – The mediating effect of digital orientation and controllers' added value in explaining organizational implications of advanced analytics," Journal of Management Control, vol. 33, pp. 403–446, 2022. https://doi.org/10.1007/s00187-022-00343-w.

[2] Hallo L. and Nguyen T., "Holistic View of Intuition and Analysis in Leadership Decision-Making and Problem-Solving," Administrative Sciences, vol. 12, no. 1, p. 4, 2022. https://doi.org/10.3390/admsci12010004.

[3] Choudhary S., "Inventory Optimization Using an AI-Powered Holistic Model for Decision-Support System," International Journal for Research in Applied Science and Engineering Technology, vol. 12, pp. 414–421, 2024. https://doi.org/10.22214/ijraset.2024.58732.

[4] Sateesh Reddy C. and Nimo J.R.A., "Analyzing the Impact of Analytics on Organizational Performance," International Journal of All Research Education and Scientific Methods, vol. 12, pp. 3383–3390, 2023. https://doi.org/10.56025/IJARESM.2023.1201243383.

[5] S.-Y. Hung, C. C. Chen, H. S. Choi, and P. Ractham, "A holistic framework to examine the impact of user, organizational and data factors on the use of big data analytics systems," Information Research, vol. 26, no. 4, 2021, https://doi.org/10.47989/irpaper915

[6] Elgendy N., Elragal A., and Päivärinta T., "Evaluating collaborative rationality-based decisions: a literature review," Procedia Computer Science, vol. 219, pp. 647–657, 2023. https://doi.org/10.1016/j.procs.2023.01.335.

[7] Kang K., "Reformulating 'Holism' in hydropower decision-making," Systems Research and Behavioral Science, vol. 37, pp. 360–368, 2020. https://doi.org/10.1002/sres.2635.

[8] Morales-Arevalo J.C. and Rodríguez C., "A Systematic Review of the Benefits and Challenges of Data Analytics in Organizational Decision-making," International Journal of Advanced Computer Science and Applications, vol. 16, 2025. https://doi.org/10.14569/IJACSA.2025.0160221.

[9] Lin J., Harada K., and Goto H., "Designing and Implementing Organizational Learning Analytics Following a Holistic Planner Tool," 2023. https://doi.org/10.2139/ssrn.4655662.

[10] Muñoz M.A., Pineda S., and Morales J.M., "A bilevel framework for decision-making under uncertainty with contextual information," Omega, vol. 108, p. 102575, 2022. https://doi.org/10.1016/j.omega.2021.102575.

[11] Kolar E., Tušak M., Robnik P., and Biloslavo R., "Strategic decisions: How good does it fit the decision-making process into the strategic management process?" Journal of Engineering Management and Competitiveness, vol. 14, pp. 99–111, 2024. https://doi.org/10.5937/JEMC2402099K.

[12] Frost D., Gericke O., Di Bari R., Balangé L., Zhang L., Blagojevic B., et al., "Holistic Quality Model and Assessment—Supporting Decision-Making towards Sustainable Construction Using the Design and Production of Graded Concrete Components as an Example," Sustainability, vol. 14, p. 11269, 2022. https://doi.org/10.3390/su141811269.

[13] Khan A., Khan S., Rahimzai A.A., and Abdullah S., "Extended version of decision-making model for industrial robot selection via fractional continuous fuzzy information," Journal of Big Data, vol. 12, p. 158, 2025. https://doi.org/10.1186/s40537-025-01190-x.

[14] Romiti A., Del Vecchio M., Cavicchi C., and Vagnoni E., "Healthcare organizations in crisis context: decision-making models and roles of CEOs," BMC Health Services Research, vol. 25, p. 273, 2025. https://doi.org/10.1186/s12913-025-12420-6.

[15] Chacoma A. and Billoni O.V., "Emergent complexity in the decision-making process of chess players," Scientific Reports, vol. 15, p. 23234, 2025. https://doi.org/10.1038/s41598-025-06335-9.

[16] Davenport T.H., Harris J., and Abney D., Competing on Analytics: The New Science of Winning. Boston: Harvard Business Review Press.

[17] Ghoshal T., Bedeley R.T., Iyer L.S., and Bhadury J., "Business Analytics Capabilities and Use: A Value Chain Perspective," in Analytics and Data Science: Advances in Research and Pedagogy, Cham: Springer, 2018, pp. 41–54. https://doi.org/10.1007/978-3-319-58097-5_5.

[18] Bedeley R.T., Ghoshal T., Iyer L.S., and Bhadury J., "Business Analytics and Organizational Value Chains: A Relational Mapping," Journal of Computer Information Systems, vol. 58, pp. 151–161, 2018. https://doi.org/10.1080/08874417.2016.1220238.

[19] Jeyaprabha B., Kumar S.R., Bolla R.L., Bhatt A.S., Sera R.J., and Arora K., "Data-Driven Decision-making in Management: Leveraging Big Data Analytics for Strategic Planning," in Proceedings of the First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies, 2025, pp. 1000–1003. https://doi.org/10.1109/CE2CT64011.2025.10939548.

[20] Borges M., Bernardino J., and Pedrosa I., "Data-driven decision-making strategies applied to marketing," in Proceedings of the 16th Iberian Conference on Information Systems and Technologies, 2021, pp. 1–7. https://doi.org/10.23919/CISTI52073.2021.9476506.

[21] Staneviciene E., Gudoniene D., Punys V., and Kukstys A., "A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning," Sustainability, vol. 16, p. 10442, 2024. https://doi.org/10.3390/su162310442.

[22] Xavier M.K., Leal G.C.L., Guerino G.C., Coleti T.A., and Balancieri R., "Automatic Complaints Classification in E-Commerce: A Case Study Using CRISP-DM," Journal on Interactive Systems, vol. 16, pp. 256–266, 2025. https://doi.org/10.5753/jis.2025.4661.

[23] Morales-Arevalo J.C., Aquise-Gonzales E.M., Carpio-Ore W.Y., Sáenz E.V.M., Mazzarri-Rodriguez C.J., Remotti-Becerra E.E., et al., "Data Mart Design to Increase Transactional Flow of Debit and Credit Card in Peruvian Bodegas," International Journal of Advanced Computer Science and Applications, vol. 16, 2025. https://doi.org/10.14569/IJACSA.2025.0160216.

[24] Yessad L. and Labiod A., "Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault," in Proceedings of the International Conference on System Reliability and Science, 2016, pp. 95–99. https://doi.org/10.1109/ICSRS.2016.7815845.

[25] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, The Data Warehouse Lifecycle Toolkit, 2nd ed. Indianapolis, IN, USA: Wiley, 2008.

[26] A. Amir, C. Vinke, and R. van Heijningen, "Horti-IoT: a new data framework for Dutch horticulture," International Food and Agribusiness Management Review, vol. 27, no. 5, pp. 858–881, 2024, https://doi.org/10.22434/IFAMR1124.

[27] Bahaa S., Ghalwash A.Z., and Harb H., "DataOps Lifecycle with a Case Study in Healthcare," International Journal of Advanced Computer Science and Applications, vol. 14, 2023. https://doi.org/10.14569/IJACSA.2023.0140115.

[28] AbouZaid A., Barclay P.J., Chrysoulas C., and Pitropakis N., "Building a modern data platform based on the data lakehouse architecture and cloud-native ecosystem," Discover Applied Sciences, vol. 7, p. 166, 2025. https://doi.org/10.1007/s42452-025-06545-w.

[29] Peffers K., Tuunanen T., Rothenberger M.A., and Chatterjee S., "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, vol. 24, pp. 45–77, 2007. https://doi.org/10.2753/MIS0742-1222240302.

[30] Alnajjar O., Atencio E., and Turmo J., "Framework for Optimizing the Construction Process: The Integration of Lean Construction, Building Information Modeling (BIM), and Emerging Technologies," Applied Sciences, vol. 15, p. 7253, 2025. https://doi.org/10.3390/app15137253.

[31] Caballero I., Gualo F., Rodríguez M., and Piattini M., "BR4DQ: A methodology for grouping business rules for data quality evaluation," Information Systems, vol. 109, p. 102058, 2022. https://doi.org/10.1016/j.is.2022.102058.

[32] Gualo F., Caballero I., Rodríguez M., and Piattini M., "A Data Quality Model for Master Data Repositories," Informatica, vol. 34, pp. 795–824, 2023. https://doi.org/10.15388/23-INFOR534.

[33] Hegselmann S., Storck M., Gessner S., Neuhaus P., Varghese J., Bruland P., et al., "Pragmatic MDR: a metadata repository with bottom-up standardization of medical metadata through reuse," BMC Medical Informatics and Decision-making, vol. 21, p. 160, 2021. https://doi.org/10.1186/s12911-021-01524-8.

[34] Hu Z., Wang A., Duan Y., Zhou J., Hu W., and Wu S., "Toward Better Semantic Interoperability of Data Element Repositories in Medicine: Analysis Study," JMIR Medical Informatics, vol. 12, p. e60293, 2024. https://doi.org/10.2196/60293.

[35] Malatji M., "Management of enterprise cyber security: A review of ISO/IEC 27001:2022," in Proceedings of the International Conference on Cyber Management and Engineering, 2023, pp. 117–122. https://doi.org/10.1109/CyMaEn57228.2023.10051114.

[36] Utomo W.H., Rosalina R., and Afriyadi A., "Privacy-preserving data mining optimization for big data analytics using deep reinforcement learning," Indonesian Journal of Electrical Engineering and Computer Science, vol. 36, pp. 1929–1937, 2024. https://doi.org/10.11591/ijeecs.v36.i3.pp1929-1937.

[37] Miller R., Chan S.H.M., Whelan H., and Gregório J., "A Comparison of Data Quality Frameworks: A Review," Big Data and Cognitive Computing, vol. 9, p. 93, 2025. https://doi.org/10.3390/bdcc9040093.

[38] Aradiel Castañeda H., Acosta De La Cruz P.R., Mas Azahuanche G.A., Geronimo Vasquez A.H., and Flores Salinas J.A., "Generating Dashboards with Python on the RPA process for optimizing the assignment of gardening tasks," in Proceedings of the 4th LACCEI International Multiconference on Entrepreneurship, Innovation and Regional Development, 2024. https://doi.org/10.18687/LEIRD2024.1.1.709.

[39] Nunes F., Alexandre E., and Gaspar P.D., "Implementing Key Performance Indicators and Designing Dashboard Solutions in an Automotive Components Company: A Case Study," Administrative Sciences, vol. 14, p. 175, 2024. https://doi.org/10.3390/admsci14080175.

[40] Komóczi M. and Kósa K., "Storytelling as narrative health promotion in community psychiatry: a quasi-experimental study," BMC Psychiatry, vol. 25, p. 376, 2025. https://doi.org/10.1186/s12888-025-06816-1.

[41] Song Y., Li W., Wang Z., and Hu Y., "Visual storytelling through the void: a quantitative analysis of empty shot distribution across film genres," Humanities and Social Sciences Communications, vol. 12, p. 232, 2025. https://doi.org/10.1057/s41599-025-04527-2.

[42] Han S.Y., Yoo J., Michels A., Kang J.-Y., Wang S., and Kim J.-S., "Vulnerable Neighborhood Explorer (VNE): An Open-Source Visual Analytics Tool for Exploring Social Vulnerability to Disasters across Different Neighborhoods," SoftwareX, vol. 31, p. 102233, 2025. https://doi.org/10.1016/j.softx.2025.102233.

[43] Yan L., Martinez-Maldonado R., Jin Y., Echeverria V., Milesi M., Fan J., et al., "The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics," Computers & Education, vol. 234, p. 105322, 2025. https://doi.org/10.1016/j.compedu.2025.105322.

[44] Simon H.A., The New Science of Management Decision. New York: Harper & Brothers, 1960.

[45] Liberatore M.J. and Wagner W.P., "Simon's Decision Phases and User Performance: An Experimental Study," Journal of Computer Information Systems, vol. 62, pp. 667–679, 2022. https://doi.org/10.1080/08874417.2021.1878476.

[46] Hevner A. and Chatterjee S., "Design Science Research in Information Systems," in Design Research in Information Systems: Theory and Practice, Boston: Springer, 2010, pp. 9–22. https://doi.org/10.1007/978-1-4419-5653-8_2.

[47] Buitrago Pulido R.D., Salinas J., and Boude Figueredo O., "Validation of a Model for the Formalization of Personal Learning Pathways Through Expert Judgment," Journal of Higher Education Theory and Practice, vol. 23, 2023. https://doi.org/10.33423/jhetp.v23i12.6283.

[48] Stander H.-M., Cohen B., Harrison S.T.L., and Broadhurst J.L., "Validity of using expert judgements to inform multiple criteria decision analysis: Selecting technologies for sulfide-enriched fine coal waste reuse," Journal of Cleaner Production, vol. 415, p. 137671, 2023. https://doi.org/10.1016/j.jclepro.2023.137671.

[49] Mikša M. and Sikirica N., "The Evolution and Applications of the Technology Acceptance Model," vol. 4, pp. 154–157, 2024. https://doi.org/10.15341/mce(2470-4180)/05-06.04.2024/005.

[50] Harrati N., Bouchrika I., Tari A., and Ladjailia A., "Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis," Computers in Human Behavior, vol. 61, pp. 463–471, 2016. https://doi.org/10.1016/j.chb.2016.03.051.

[51] Guillemot J., Ortiz A., Valdivieso E., Muñoz G., and Cisneros-Heredia D.F., "Approach for participatory Delphi method for structuring voices, population empowerment and research training," MethodsX, vol. 13, p. 103019, 2024. https://doi.org/10.1016/j.mex.2024.103019.

[52] Lynn M. R., "Determination and quantification of content validity," Nursing Research, vol. 35, no. 6, pp. 382–385, 1986. https://doi.org/10.1097/00006199-198611000-00017

[53] Robles Garrote P. y Rojas M. C., "La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada," Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas, vol. 9, no. 18, pp. 124–139, 2015. https://doi.org/10.26378/rnlael918259

Multi-Scale Curvelet-Based Directional Denoising for Chest X-Ray Images

Neenu Sebastian*, B. Ankayarkanni
Department of Computer Science and Engineering-School of Computing,
Sathyabama Institute of Science and Technology (Deemed to be University), Chennai 600119, India

***Abstract*—In modern healthcare, medical imaging has a significant role in understanding the structure and functioning of the human body, which helps doctors to diagnose, to plan the treatment, and to monitor the disease. Chest X-rays are widely used for the early detection and treatment of various lung infections. The effectiveness and the accuracy of the diagnosis depend largely on the quality of the medical images. Chest X-rays often suffer from Gaussian and Poisson noise, which affects the visibility of fine anatomical structures. Although methods like Gaussian filtering, NLM, and GAN have been used, they often compromise between denoising and retaining edge details. A robust denoising algorithm, Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD), is proposed to denoise medical chest X-ray images, which uses the curvelet transform coefficients. The performance of the proposed MCF-DD model was evaluated on the Chest X-Ray Images dataset from the Kaggle repository and DICOM images from the MIDRC-RICORD-1C dataset. MCF-DD achieved a PSNR of 36.57dB and SSIM of 0.9062 on Kaggle images, and 40dB PSNR with 0.9412 SSIM on DICOM images, indicating strong denoising performance across both datasets.**

***Keywords*—Image denoising; medical image; Chest X-Ray; Poisson noise; Gaussian noise; Curvelet Transform**

I. Introduction

Medical images are used in healthcare for the diagnosis and monitoring of human diseases and health. Chest X-rays are of significant relevance in current medical diagnosis applications, as they are the most widely used cost-effective modality for diagnosing lung infections such as pneumonia, lung cancer, and COVID-19 [1]. The quality of chest X-rays is often compromised by the presence of noises such as Poisson and Gaussian noise [2]. The main sources of noise in X-ray images are scattered radiations, fluctuations in electronics, and quantum noise. To reduce the exposure of patients to ionizing radiation, low-dose imaging protocols that prescribe the usage of low photon count are used during medical imaging. Due to statistical variations in photon counts, Poisson noise gets added to X-ray images. As a result of these noises, the pixel distribution of the images will be altered, and this may result in the loss of image features such as texture, edges, intensity, etc. The presence of this noise affects the exact diagnosis and monitoring of the disease [3]. Enhancing the quality of Chest X-Ray images through effective denoising techniques is essential to support clinical procedures, as demonstrated by image-guided interventions that rely on clear imaging for precise treatment planning. So, image denoising is a required pre-processing stage for applications using medical images and so it is crucial to develop effective denoising techniques. Also, reliable denoising of chest X-rays under mixed Poisson and Gaussian noise remains a key challenge. Existing methods often fail to effectively suppress noise while preserving fine anatomical structures that are critical for accurate pneumonia diagnosis.

The various noise reduction methods can be broadly classified into four categories: conventional filtering, transform domain filtering, deep learning-based methods [4], and hybrid models [5]. For denoising, the conventional filtering methods use linear filters such as a Gaussian filter, mean filter, etc., and nonlinear filters such as a median filter [6], bilateral filter, etc. These filtering methods have low computational complexity, but the denoised images will be blurred, resulting in low resolution. In transform domain filtering, different transform functions are used to remove image noise. The advantages of transform methods include edge preservation and multi-resolution analysis, but after denoising using transform methods, the resultant image may contain artifacts that occur as a result of poorly chosen parameters and threshold values. Deep learning-based denoising methods use neural network architectures, particularly convolutional neural networks, autoencoders, and generative models, to learn complex noise patterns directly from data [7]. These approaches can adapt to various noise distributions and often outperform traditional techniques in preserving fine structural details. However, their performance heavily depends on the availability of large, high-quality annotated datasets, and they typically require significant computational resources for training and inference. Hybrid denoising models integrate multiple techniques, typically combining conventional or transform domain methods with deep learning approaches to exploit the strengths of each technique. These models aim to enhance denoising accuracy and robustness across diverse noise types. Recent studies report that deep learning methods show poor generalization across scanners, dose levels, and patient subgroups, and may cause over-smoothing or hallucinated textures in chest X-rays.

The above-discussed methods exhibit limitations such as blurring in conventional filters to data dependency in deep learning models. So, there is a need for an adaptive and structure preserving denoising framework for Chest X-Ray images. To address these challenges, Multi Scale Curvelet Filtering with Directional Denoising (MCF DD) framework is proposed. MCF DD employs curvelet based dynamic thresholding, which adaptively suppresses noise while preserving edge and texture details across scales and directions.

*Corresponding author

To overcome the lack of orientation-sensitivity in traditional methods, Gabor and Morlet filters are integrated to enhance directional and frequency-specific features. The inclusion of an anisotropic Gaussian filter ensures edge-preserving smoothing, minimizing the risk of blurring diagnostic boundaries. Unlike deep learning models, MCF DD does not rely on large annotated datasets or intensive training, offering a computationally efficient solution. Furthermore, by performing coefficient-level fusion of multiple filtered outputs, MCF DD combines the strengths of each component filter, achieving superior denoising performance without the architectural complexity of conventional hybrid models.

Unlike existing methods that rely on fixed global filtering, the proposed MCF-DD uses patch-wise adaptive noise characterization to guide noise-specific directional curvelet denoising, enabling effective handling of spatially varying mixed noise. In the proposed approach, a patch transformation method is applied where the input image patches are obtained by segmenting the input image, which is then fed as an input to the proposed robust denoising algorithm, Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD). In this algorithm, to capture the different anatomical structures, a multi-scale representation is needed, which can be achieved by using Curvelet Transform. By using Curvelet Transform, the input patches are divided into different scales and orientations, which helps to get the information regarding the local and global structure. To preserve the edge information, Curvelet Transform is integrated with directional filtering techniques: Gabor, Morlet, and Anisotropic Gaussian filters. This approach aims to effectively remove noise such as Gaussian and Poisson while preserving critical anatomical structures and features across multiple scales and orientations.

In summary, the key contributions of the proposed framework can be outlined as:

- A novel multi-scale denoising approach (MCF-DD) is proposed, which integrates curvelet-based multi-resolution decomposition with directional filtering to preserve structural details and suppress diverse noise types in chest X-rays.
- An initial patch-based classification mechanism is introduced that adaptively processes regions with varying noise intensities, reducing unnecessary computation and enhancing local feature preservation.
- For experimental validation, the model was tested on JPEG format chest radiographs from the Kaggle Chest X-ray dataset [8] and DICOM images from the MIDRC-RICORD-1C dataset [9]. The inclusion of both compressed and uncompressed formats ensures robustness of the proposed method across different imaging standards.

II. Related Work

Image denoising refers to the methods used to find the clear image from the noise image, as depicted in Fig. 1. The aim of denoising methods is to remove the noise from medical images and also to preserve the clinical details. According to the evaluation criteria outlined in [10], effective image denoising should maintain uniformity in flat regions, preserve edge sharpness without blurring, retain structural and textural details, keep the original contrast intact, and avoid introducing any unnatural artifacts.

In medical imaging, the different sources of noise can be low light, sensor characteristics, dust particles in scanning machines, disturbance in transmission channels, etc. The predominant types of noise encountered in medical images include Gaussian noise and Poisson noise. Gaussian noise is a type of noise that follows a Gaussian distribution [11]. Gaussian noise can be caused by different sources, such as sensor temperature fluctuations, low radiation dose, etc. Poisson noise is caused by the photon emission in Chest X-Rays. The variance of Poisson noise depends on image signal intensity. Chest X-Ray images and CT images usually contain a combination of Gaussian and Poisson noise. Various methods have been suggested to accomplish these denoising goals, ranging from traditional image processing methods to modern machine learning approaches. These methods are broadly categorized into conventional filtering, transform-based techniques, deep learning approaches, and hybrid models.

A detailed study of the application of filtering algorithms for denoising images is done in [12]. It is the most widely used denoising method. Filtering methods use mathematical models to remove noise from images. The linear filter works by computing the weighted sum of pixel values in a defined neighborhood. Gaussian[13] and mean filters are examples of linear filters that will remove the noise, but result in blurred denoised images. Median filter [14], Bilateral filter [15], and non-local means filter [16] use nonlinear adaptive filtering techniques to remove noises from images. They are better in preserving the edge information than the linear filters. In [17], advanced filtering techniques such as an adaptive Wiener filter are proposed, which generate less blurred denoised images compared to the other methods. An iterative process to reduce noise is proposed in [18], which results in better edge preservation while maintaining image quality.

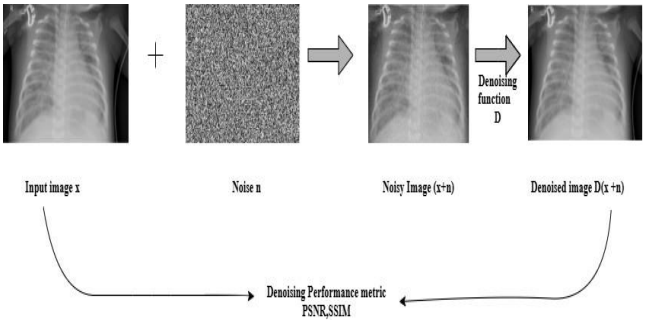


Fig. 1. Medical image denoising

Gaussian and Poisson noises are common types of noise affecting medical images. Filtering approaches work well for a particular type of noise, and their performance degrades for mixed noise [19]. So, there is a need for a denoising method capable of handling multiple types of noise.

In transform-based methods, to separate the noise and signal components, the image is converted from the spatial model to the mathematical model, where the noise and signals

have different frequency components. To suppress the frequency components of the noise, filtering or threshold methods are applied which preserve significant image structures. In Wavelet Methods [20], the wavelet threshold method was proposed, in which wavelet decomposition is used to separate noise from image signals, enabling adaptive noise removal with minimal loss of information. This method resulted in the over-smoothening of important anatomical features, as it captures only point singularity information rather than complex structural information. Adaptive Bayes Shrink thresholding [21] was proposed, in which the noise variance is calculated in each sub-band using the Bayesian estimation, which is used for the calculation of the optimal threshold value, which varies according to the noise level present in each band. DWT is an inconsistent noise removal method due to its shift variant property, in which a small shift in the image information leads to the generation of different wavelet coefficients. In [22], Stationary Wavelet Transform (SWT), which addresses the shift variant property of DWT, is proposed by eliminating down-sampling, which is done by using the unsampled filter coefficients and convolution. The increased computational cost and redundancy make it less suitable for real-time processing of medical images. The other transform-based methods are Curvelet Transform (CT) [23], [24], Contourlet Transform (CTT) [25], Shearlet Transform (ST) [26]. Transform-based methods handle Gaussian noise, but they have limitations when handling Poisson noise due to the signal's dependent nature.

Deep learning methods are widely used for medical image denoising due to their capability for automatic feature extraction. Specific noise models, which are used in traditional methods, limit their ability to adapt to mixed noise input images, while deep learning approaches learn the noise patterns from noisy images, which are later used for noise removal. A deep CNN model for handling the Gaussian noise using residual learning is proposed in [27]. A self-supervised learning method that removes Poisson noise in CT images was proposed in [28]. The requirement of multiple noisy images of the same patient challenged the efficiency of this method in medical imaging. Autoencoder-based approaches were proposed in the work [29]. In [30], an efficient deep learning model to estimate a residual image from the input image with the Gaussian noise was proposed. In GAN-based methods, an image-to-image translation approach for CT images was proposed in [31], [32]. The lack of paired training data, overfitting on small datasets, and high computational complexity challenge the application of deep learning methods for medical image denoising.

For achieving better results, several researchers have proposed hybrid approaches, in which different denoising methods with different strengths are combined together. In [33], BM3D (Block Matching and 3D filtering) and wavelet transforms have been combined together which yields better textural preservation and fine details. But this method demanded a larger memory space requirement due to 3D filtering in BM3D. In [34], the CNN's ability to learn complex patterns is combined with wavelets' multi-scale representation. Several other hybrid models have been discussed in the work [35]. Hybrid methods give better results than single method, but the need of extensive training data, need of parameter tuning, and increased computational complexity limits its application. Inspired by these approaches, the proposed work aims to improve Chest X-Ray denoising by preserving critical anatomical structures and enhancing image quality. Though there are significant developments in medical image denoising techniques, there are several challenges that are to be addressed. One major challenge is the lack of a denoising technique that can effectively handle the mixed noise types that are often present in real-world medical imaging. Another challenge is the high computational complexity of particularly deep learning and hybrid methods, which will make them less practical for medical applications. There exist numerous studies for denoising MRI and CT images, and there is a noticeable lack of research in denoising X-Ray images. Among the various medical image modalities available, this research work focuses on Chest X-Rays, which are used for the detection and diagnosis of chest infections. This study presents a novel denoising framework, Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD), specifically designed to address critical gaps in existing medical image denoising techniques. Unlike conventional, deep learning, or hybrid approaches, the proposed method effectively handles mixed noise types, which commonly affect real-world chest X-rays. By focusing on the underexplored modality of chest radiographs and prioritizing both effective noise suppression and structural detail preservation, MCF-DD introduces a practical and domain-specific solution for enhancing diagnostic quality in resource-constrained clinical settings. Recent diffusion-aware frameworks, such as DIFCNet, utilize diffusion features or reconstruction inconsistencies learned through deep generative models to characterize image degradations. In contrast, the proposed MCF-DD framework is a non-learning, noise-adaptive method that directly targets mixed Poisson Gaussian noise through patch-wise statistical characterization and directional curvelet domain filtering, making it computationally efficient.

III. Proposed Methodology

The overall workflow of the proposed Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD) framework is illustrated in Fig. 2. For denoising medical images, a new method is proposed by leveraging a multi-scale Curvelet Transform integrated with directional filtering techniques such as Gabor, Morlet, and Anisotropic Gaussian filters. This approach removes both Gaussian and Poisson noise, while preserving critical anatomical structures and features across multiple scales and orientations. Poisson noise is signal-dependent and arises from the statistical nature of X-ray photon detection, making it more prominent in low-dose imaging, where fewer photons reach the detector. In contrast, Gaussian noise is signal-independent and typically originates from electronic imperfections in imaging systems. While Poisson noise affects areas with low photon counts, Gaussian noise is uniformly distributed across the image. Both types of noise degrade image quality by reducing contrast and obscuring anatomical details. Denoising methods like wavelet transforms, CNN-based approaches, and filtering techniques are often employed to mitigate these noise effects.

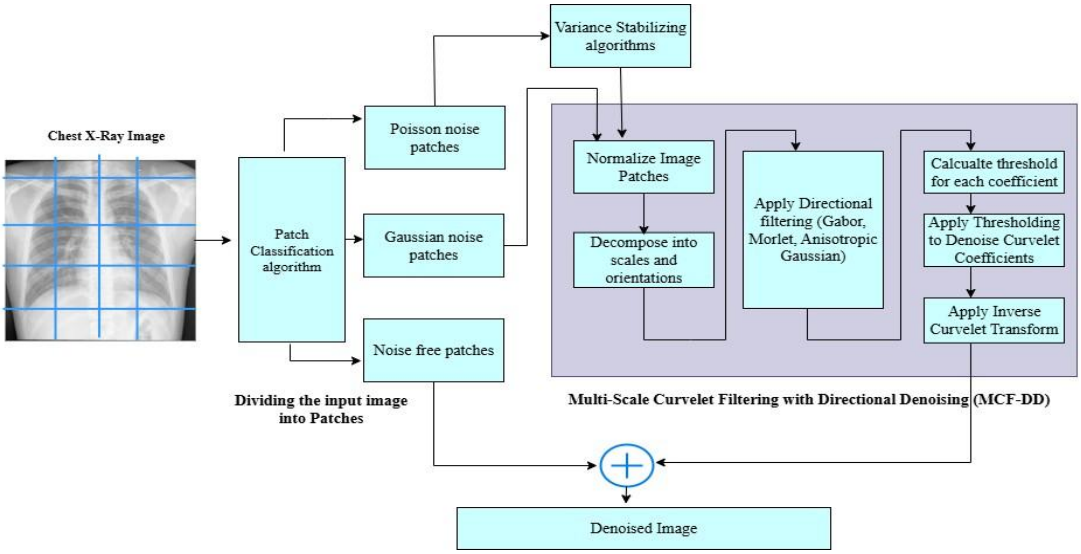


Fig. 2. System architecture

A. Patch Classification

In medical X-rays, preserving and enhancing edges and fine anatomical details is crucial for accurate diagnosis and interpretation. Conventional denoising and feature extraction methods, such as Gaussian filtering or basic wavelet transforms, often struggle to maintain these important features. The Curvelet Transform, on the other hand, represents images in terms of localized, oriented, and scale-specific components, which align better with the curvilinear structures typically present in biological tissues. To preserve and enhance subtle edges and fine anatomical details, the input medical image is divided into different patches of size 64 X 64. Each patch is analyzed to find the type of noise dominant in that patch, and correspondingly, the patch is classified into the following classes, such as the Gaussian class, Poisson class or noise-free image patch class. The noise-free image patches are preserved without any modifications and processing. The Poisson class patches will undergo Variance Stabilizing Transformations (VST) to transform the Poisson noise into a form where the variance becomes more uniform. The stabilized Poisson patch and the Gaussian noise patch will be given as input to Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD) for noise removal from the noisy image patches. The denoised image patches are aligned along with the noisy free patches to reconstruct the denoised image.

The medical image is divided into smaller, non-overlapping patches to analyze and process localized regions independently. This approach allows for identifying noise characteristics and applying appropriate denoising techniques at a granular level. This patching method enhances the ability to denoise the image efficiently while maintaining critical anatomical features. Each input image is divided into non-overlapping square patches of size 64 × 64 pixels. Prior to division, image dimensions are checked to ensure compatibility with the patch size. If the height or width is not perfectly divisible by 64, zero padding is applied to the bottom or right edges of the image. Each patch is analyzed for the dominant noise present in it. The difference in the characteristics of Gaussian and Poisson noises is used to calculate the dominant noise in the patches. Gaussian noises are uniform across all intensities. Gaussian noises will have high variance but are independent of the mean intensity. On the other hand, for Poisson noise, its variance increases proportionally to the mean intensity. Noise-free patches will have low variance-to-mean ratio. For an image patch with ‘n’ no pixels and if ‘li’ is the intensity of pixel i, the mean values of the patch ’m’ can be calculated using Eq. (1) [36]:

$$m = \frac{1}{n}\sum_{i=1}^{n} l_i \quad (1)$$

The measure of the spread of pixel intensities around the mean value gives the variance value ‘v’ of the patch which can be calculated using Eq. (2) [37]:

$$v = \frac{1}{n}\sum_{i=1}^{n}(l_i - m_i)^2 \quad (2)$$

Variance to Mean ratio (v_to_m_ratio) for image patch can be calculated using Eq. (3) [37]:

$$v_{to_m_ratio} = \frac{v}{m} \quad (3)$$

For classifying the image patches into Gaussian, Poisson and clean image clusters, Gaussian Mixture Model is used, which represents the data as mixture of multiple Gaussian distributions. In noisy Chest X-Ray image, the variance to mean ratio for image patches follows a multi modal distribution as the variance to mean ratio value’s characteristics changes depending on the type of noise dominant in that patch. GMM (Gaussian Mixture Model) is used to understand the data distribution and to determine the threshold values for classification. These threshold values dynamically vary depending on the data distribution characteristics. To fit a mixture of Gaussian distributions to variance to mean ratio, GMM initialize three Gaussian components for clean patches, Poisson patches and Gaussian patches, respectively. Then it uses Expectation Maximization algorithm [38] to adjust and improve the parameters such as mean (μ), variance (σ2), and weight (w). The mean values of the established three Gaussian distribution indicate the typical values for clean patches,

Poisson noise patches, and Gaussian noise patches, respectively. Algorithm 1 presents the patch classification.

Algorithm 1: Patch Classification Algorithm

Normalize the input image by scaling the pixel values to [0,1].

Extract *p_size* X *p_size* patches from the image.
Calculate the variance-to-mean ratio for each image patch.
Determine the dynamic thresholds T1, T2 using Gaussian Mixture Model (GMM), by fitting a 3 component GMM to the V/M ratios and use cluster means to set T1, T2.
For all image patches do
Compute mean and variance for the patch.
Compute *v_to_m_ratio.*
If *v_to_m_ratio*≥*T*2then
Assign to Gaussian_noise_patches
else if *T*1 ≤ *v_to_m_ratio* ≤ *T*2 then
Assign to Poisson_noise_patches
Else
Assign to Clean_patches
end if.
End

The thresholds T1 and T2 are then calculated as the midpoints between the adjacent Gaussian means to establish decision boundaries between the noise classes. T1 is set as the average of the mean values of the clean and Poisson noise distributions, and T2 is calculated as the average of the Poisson and Gaussian noise distributions. This adaptive thresholding approach ensures that the classification boundaries are not arbitrarily chosen but instead dynamically adapt to the distribution of noise in the dataset. The patches whose variance-to-mean ratio is greater than T2 are classified as Gaussian patches, and the patches with their variance-to-mean ratio between T1 and T2 are classified as Poisson patches. Patches with a variance-to-mean ratio less than T1 are classified as clean patches. Unlike a fixed threshold, GMM-based thresholds account for variations in image intensity ranges and different noise levels, making the method more robust and dataset independent.

The Poisson noise patches are transformed to Gaussian noise characteristics by applying a variance stabilizing technique called Anscombe transformation [39], which can be mathematically represented using the following Eq. (4) [39], where 'X' is the Poisson noise and 'Y' is the transformed noise signal, which resembles Gaussian noise, as the variance is stabilized.

$$Y = 2\sqrt{X + \frac{3}{8}} \tag{4}$$

To qualitatively support the GMM-based noise categorization, visual examples of image patches exhibiting different noise characteristics are displayed in Fig. 3. These patches, extracted from Chest X-Rays, represent samples categorized into distinct Gaussian components by the GMM. Each cluster corresponds to a different noise, such as clean patches, Gaussian patches and Poisson patches. The visual distinction in texture and granularity supports the adaptive noise categorization capability of the proposed algorithm.

B. Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD)

Medical images contain anatomical structures of varying sizes, like small cellular structures to large-scale structures like organ boundaries. In Chest X-Ray images, structures such as the lung trachea, Primary Bronchi, Bronchial Tree, and lungs constitute the larger structures, and small blood vessels, Alveolar Sacs, Bronchioles, and Capillaries are examples of smaller structures. So, it's necessary to have a multi-scale approach that retains and preserves both the small-scale and large-scale features. Also, the noise in medical images can be present at different scales, so a multi-scale approach is proposed, which addresses noise at each scale independently.

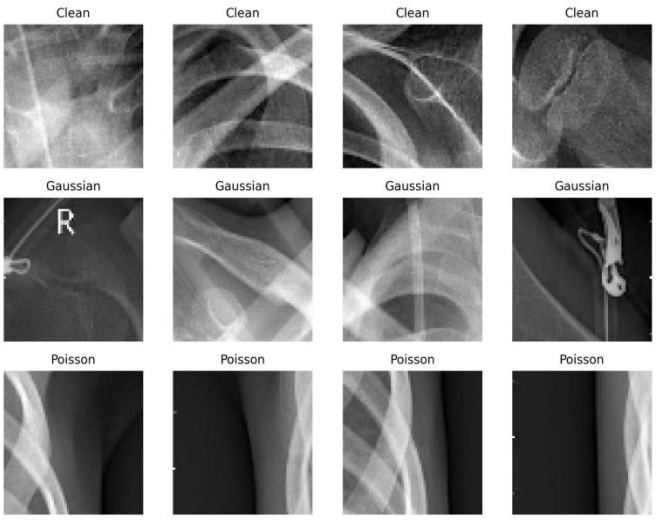


Fig. 3. Representative image patches from different Gaussian components identified by the GMM. Each row shows patches associated with a distinct noise category.

In Curvelet Transform, the images are decomposed into multiple scales and orientations, which captures both the small-scale and large-scale information in the X-Ray image. The Curvelet Transform outperforms the other mathematical transforms in capturing edges, curves, and fine details in images, which is particularly useful for medical imaging. It offers better edge preservation compared to other transforms. Algorithm 2 presents the MCF-DD Denoising algorithm. The different steps in Curvelet Transform are the following:

1) Decomposition: The Curvelet Transform decomposes the image into several sub-bands. Each sub-band captures the information in a specific orientation and scale. As a result of decomposition, the image is divided into multiple scales, where each scale corresponds to a different level of detail. Each scale contains multiple orientations that capture the edge and curve information in different directions like horizontal, vertical and diagonal. The mathematical representation of decomposition can be represented by Eq. (5), in which N (X, Y) the noisy image patch which is decomposes to 'S' number of scales with 'O' Orientations in each scale.

$$N(X,Y) \approx \sum_{s=1}^{S} \sum_{o=1}^{O} C_{s,o}(X,Y) \tag{5}$$

Algorithm 2: MCF-DD Denoising Algorithm

Apply Variance Stabilizing Transform (VST) to convert Poisson noise to a Gaussian-like distribution:

$$Y' = T(Y)$$

Perform Curvelet Decomposition on Y' using Fast Discrete Curvelet Transform (FDCT):

$$\Theta_Y = C(Y')$$

for all curvelet sub-bands $\theta_i \in \Theta_Y$ do

Apply directional filters (Gabor, Morlet wavelet, Anisotropic Gaussian) to θ_i

Estimate adaptive threshold τ_i using local statistics.

Apply thresholding to obtain denoised coefficients:

$$\hat{\theta}_i = F(\theta_i, \tau_i)$$

Reconstruct the image from denoised coefficients:

$$\hat{Y'} = C^{-1}(\hat{\Theta}_Y)$$

Apply inverse variance stabilizing transform to get the final denoised image:

$$\hat{X} = T^{-1}(\hat{Y'})$$

return $\hat{X}$

End

2) Apply directional filtering to each sub-band: Directional filtering is applied to each sub-band, in different scales and orientations, to capture localized curve and edge details. In chest X-Ray images, as both the fine and larger details have to be preserved, a combination of Gabor filter, Morlet filter, and Anisotropic Gaussian filters is used. This combination of filters provides the best trade-off between accuracy, detail preservation, and computational efficiency.

a) Gabor filter: Gabor filter [40], [41] captures the orientation and scale-specific information like edges, textures, and curved structures in images. Multi-orientation Gabor filters with orientations at 0°, 45°, 90°, and 135°, and a fixed frequency of 0.25 cycles/pixel are applied. They are used for texture analysis and feature extraction. Gabor filters are mathematically described by the following Eq. (6) [41]:

$$G(x,y) = \exp\left(-\frac{x^2+\gamma^2y^2}{2\sigma^2}\right)\cos\left(2\pi f \cdot (x\cos\theta + y\sin\theta)\right) \quad (6)$$

where,

θ - Orientation of the Gabor filter.

σ -Gaussian envelope's standard deviation.

f- Frequency of the sinusoidal component.

b) Morlet filter: Morlet filter [42] is used to capture localized frequency and orientation information. It combines a Gaussian window with a sinusoidal oscillation. They provide scale and orientation selectivity and so they are preferred for detecting features at multiple scales. Standard 2D complex Morlet wavelets were applied centred around a spatial frequency of 0.8 with a bandwidth parameter of 1. Morlet filters are mathematically described by the following Eq. (7) [42]:

$$\psi(x,y) = \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \cdot \exp(j2\pi f_0 x) \quad (7)$$

where,

σ - Gaussian envelope's standard deviation.

f0 - Wavelet Central frequency.

j - Imaginary unit (j2= −1).

x, y - Spatial coordinates.

c) Anisotropic Gaussian filter: Anisotropic Gaussian filter [43] is used for directional smoothing by adjusting the standard deviations in different axes. These filters are better at maintaining anatomical boundaries and ensuring noise reduction without losing critical details. These were applied with a variable standard deviation across directions to enhance the preservation of directional edges as given by Eq. (8) [43]:

$$G(x,y) = exp\left(-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right)\right) \quad (8)$$

where,

Σx - X-direction Standard Deviation

Σy - Y-direction Standard Deviation

(x, Y) - Spatial Coordinates

3) Threshold calculation for denoising in curvelet transform: After obtaining the Curvelet coefficients, a threshold value for each coefficient to suppress noise is calculated. The noise thresholding technique for denoising was implemented using the median absolute deviation (MAD) method, as proposed in [44]. This method is effective for estimating noise variance in the transformed domain. Specifically, the noise level $\hat{\sigma}$ is estimated from the finest scale of Curvelet coefficients using Eq. (9) [44]:

$$\hat{\sigma} = \frac{\text{median}(|C_{i,j}|)}{0.6745} \quad (9)$$

where, $C_{i,j}$ are the Curvelet coefficients at the highest frequency scale. A universal threshold is then computed using Eq. (10) [44]:

$$T = \hat{\sigma}\sqrt{2\log N} \quad (10)$$

With N being the number of coefficients in the sub-band. Hard thresholding is applied by comparing the squared magnitude of each coefficient to the threshold. If $|C_{i,j}|^2 < T^2$, the coefficient is set to zero; otherwise, it is retained as given in Eq. (11) [44]:

$$\widehat{C_{i,j}} = \begin{cases} 0, if\left|C_{i,j}\right|^2 < T^2 \\ C_{i,j}, Otherwise \end{cases} \quad (11)$$

This operation removes insignificant coefficients likely representing noise while preserving meaningful structural details. The denoised image is then reconstructed using the inverse Curvelet Transform.

4) Reconstruction of the image: After denoising, reconstruction of the image is done by performing the Inverse Curvelet Transform, as given by Eq. (12) [45]. The denoised image patch is created back from the modified Curvelet

coefficients. The inverse transform aggregates all the scales and orientations back into the original pixel grid while retaining the geometrical structures. The reconstruction ensures that both edges and curves remain intact and the essential features are preserved.

$$\hat{f}(x,y) = C^{-1} \sum_s \sum_\theta \sum_k c_{s,\theta,k} \psi_{s,\theta,k}(x,y) \quad (12)$$

where,

C−1 is the inverse Curvelet Transform.

cs,θ,k is the Curvelet coefficients at scale s, orientation θ, and position k.

ψs,θ,k(x, y) is the Curvelet basis functions.

The summation is performed over all scales s, orientations θ, and positions k.

C. Patch Reconstruction

An empty image with the same size as original image is initialized. A weight matrix is maintained for keeping track of the number of patches contributing to each pixel. The denoised patches are placed in their original locations. The weight matrix is updated to normalize the overlapping regions. The reconstructed image is normalized using the weight matrix to maintain the consistency in intensities.

D. Parameter Settings and Selection Principles

The key parameters used in the various stages of the proposed Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD) framework are listed in Table I. The selection of each parameter was guided by a combination of empirical validation, literature references, and optimization trials to ensure robustness and optimal denoising performance. Gabor and Morlet parameters were chosen to ensure effective multi scale texture representation, while anisotropic Gaussian values were fixed to balance noise suppression and edge preservation. The dynamic thresholds were computed adaptively from the statistical distribution of curvelet coefficients, ensuring data-driven noise separation reproducibility. The curvelet transform was implemented using the wrapping-based discrete formulation, and sub-band coefficients were processed with a GMM-based adaptive thresholding scheme. Gabor, Morlet, and anisotropic Gaussian filters were applied using the frequencies and bandwidths listed, and their outputs were fused by weighted averaging based on local coefficient energy. The fused coefficients were then inverted through the curvelet synthesis step to obtain the final denoised image.

TABLE I. KEY PARAMETER SETTINGS USED IN MCF-DD

Component	Parameter	Value	Justification
Curvelet Transform [46]	Number of Scales (S)	4	Captures multi-scale features with balance between detail and computational cost.
Curvelet Transform [47]	Number of Orientations (N_θ)	8	Provides directional sensitivity while avoiding redundancy.
Gabor Filter [40]	Frequency	0.6	Optimized for texture preservation in noisy lung regions.
Morlet Filter [42]	Sigma	1.5	Balances frequency and spatial localization.
Anisotropic Gaussian [43]	σx, σy	(1.2, 2.0)	Emphasizes edge preserving smoothing.
Thresholding [47]	Adaptation Method	GMM based dynamic	Adapts to local statistics and noise intensity using data driven classification.

IV. RESULTS AND DISCUSSION

A. Patch-Based Classification Performance

To evaluate the performance of patch classification approach for the noise type classification, experiments were conducted on Chest X-Ray images from Kaggle dataset [8]. The evaluation is done for the categorization of Gaussian noise, Poisson noise and also for finding the noise free patches. The experiments were conducted by dividing the input image into 32 X 32 patches and 64 x 64 patches. A synthetic noise generation strategy was employed to create patch level ground truth labels, as no publicly available Chest X-Ray dataset includes annotations identifying the dominant noise type at the patch-level.

A set of high-quality Chest X-Ray images with minimal visible noise was selected as the baseline reference, serving as the clean images for controlled noise synthesis. Gaussian noise was added using predefined mean and variance parameters, while Poisson noise was introduced based on fixed photon intensity distributions. A subset of patches was intentionally left unaltered to represent clean regions. Because each noise type was assigned deterministically during the synthesis process, precise and unambiguous ground truth labels could be maintained for every patch. The results were plotted in Fig. 4 and Patch classification is calculated by using Eq. (13) [48]:

$$\text{Accuracy} = \frac{\text{Correctly Classified Patches}}{\text{Total Patches}} \times 100\% \quad (13)$$

From Fig. 4, we can infer that larger patches improves classification accuracy. This suggests that larger patches retain more structural and statistical information, making it easier to distinguish noise patterns. The improvement in accuracy with 64×64 patches highlight their effectiveness in preserving spatial features critical for classification.

B. MCF-DD Performance

The experimental results demonstrate the effectiveness of the proposed denoising technique, Multi-Scale Curvelet Filtering with Directional Denoising in comparison to selected denoising methods.

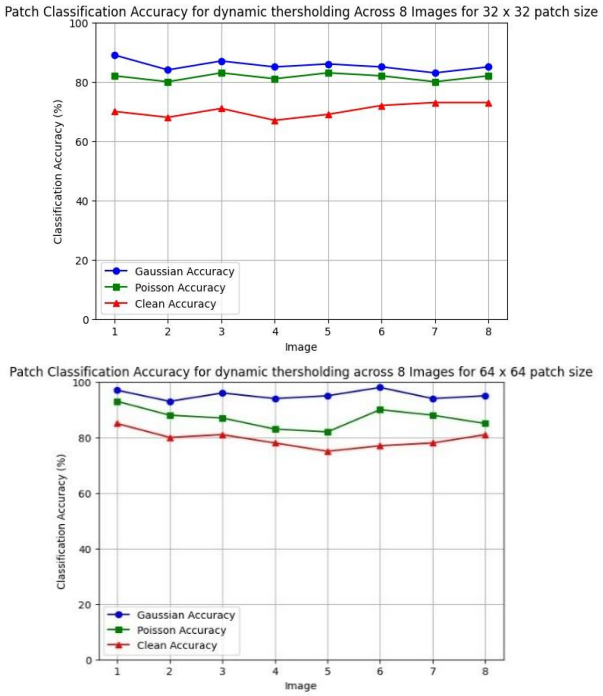


Fig. 4. Effect of patch size on classification accuracy.

Experimental study was performed on images from the Chest X-Ray Images dataset collected from the Kaggle repository [8], in which the chest radiographs are present in JPEG format. To analyze the impact of compression artifacts experimental study was also performed on the DICOM format CXR images from the Medical Imaging Data Resource Center (MIDRC) – RSNA International COVID Radiology Database (RICORD) Release 1c [9]. The MIDRC-RICORD-1C dataset provides high quality DICOM images, allowing for a more comprehensive evaluation of the denoising performance on uncompressed medical images.

To evaluate the robustness of the proposed denoising methodology, synthetic noise was added to the clean Chest X-Ray images. Specifically, Gaussian noise with a pre-defined mean of zero and variance σ2 = 0.01 was applied to simulate additive noise commonly encountered in imaging systems. Additionally, Poisson noise was introduced based on the intensity distribution of each image to replicate the photon counting noise inherent in X-ray acquisition processes. These noise models allow us to simulate realistic degradation scenarios for comprehensive evaluation. All denoisers were executed under identical noise conditions and preprocessing. Classical baselines followed standard reference implementations such as BM3D with σ matched to the corruption level, NLM with a 7×7 patch and 21×21 search window, and Wavelet denoising using soft threshold $\lambda = 3\sigma$. DnCNN was evaluated using its publicly released pretrained Gaussian model with inputs normalized to [0,1]. GAN-based baselines were retrained on the same noisy clean pairs used in the proposed work, by using U Net generator, Patch GAN discriminator, Adam optimizer (LR = 2×10^{-4}), and 100 training epochs with 256×256 patches. All methods were run on identical hardware and image sizes for consistent comparison. PSNR and SSIM are used in this study to assess denoising performance at the pixel and structural levels, and diagnostic evaluation of the proposed framework has been reported previously [49].

1) Peak Signal-to-Noise Ratio (PSNR): This metric is used to measure the quality of a denoised image compared to the original image, as shown in Eq. (14) [48]. The quality of the denoised image increases as the PSNR value increases.

$$PSNR = 10log_{10}(M^2/MSE) \quad (14)$$

2) Structural Similarity Index (SSIM): This metric is used to calculate the structural similarity between the denoised image and the original image. It is calculated by comparing different image features such as pixelpatterns, luminance, pixel intensities, etc as given in Eq. (15) [37].

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \quad (15)$$

where, x and y are the two images being compared. The constants C1 and C2 used in the SSIM calculation follow standard values as originally proposed in [37], typically set as $C1 = (K_1L)^2$ and $C2 = (K_2L)^2$, where $K_1 = 0.01$, $K_2 = 0.03$, and L is the dynamic range of pixel values.

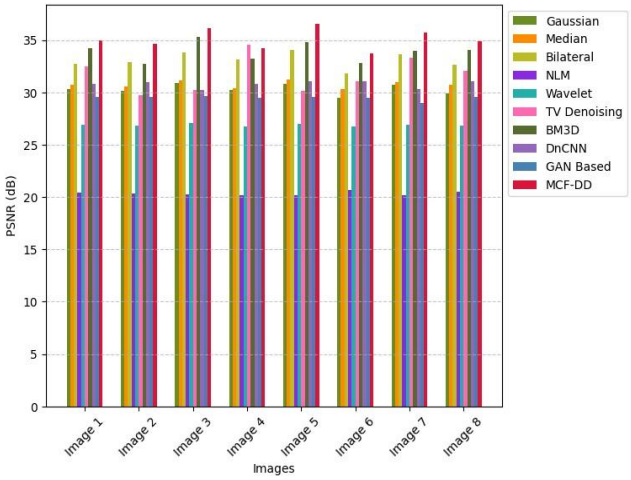


Fig. 5. PSNR comparison of denoising approaches with the proposed method, MCF-DD [Chest X-Ray Dataset].

In Fig. 5, the PSNR value obtained by the proposed MCF-DD method and the other state-of-the-art methods for denoising images from the chest X-ray data set are plotted. The results indicate that the proposed method yields a higher PSNR rate compared to the other existing methods. In Table II, the comparison of the proposed method to the other methods using the images from Chest X-Ray Dataset is given. The lower performance of NLM is mainly due to oversmoothing of fine anatomical details and reduced effectiveness in handling mixed noise, leading to lower PSNR values compared with the proposed transform-based framework. For deep learning-based denoising methods, the dataset was divided into training, validation, and testing subsets with proportions of 70%,15%, and 15%, respectively, using stratified random sampling to maintain class balance. To prevent data leakage, images from the same patient were included in only one of these subsets.

Synthetic noisy images were generated independently by adding the same noise models (Gaussian noise with zero mean and variance σ2 = 0.01, and Poisson noise based on image intensity) to the clean images in each subset. The deep learning models were trained exclusively on the noisy images from the training set, using the corresponding clean images as ground truth. Evaluation was performed on the noisy images in the testing set, ensuring no overlap or exposure of test data during training. This protocol guarantees a fair and unbiased comparison between deep learning and traditional denoising methods under consistent noise conditions.

Similarly, in Fig. 6, the SSIM values of the denoised image obtained by the proposed method and other techniques using images from the Chest X-Ray Dataset are represented. The higher the SSIM value index, the better the performance in preserving the structure of the denoised image. The proposed method has the highest SSIM values when compared with the SSIM values obtained by other methods.

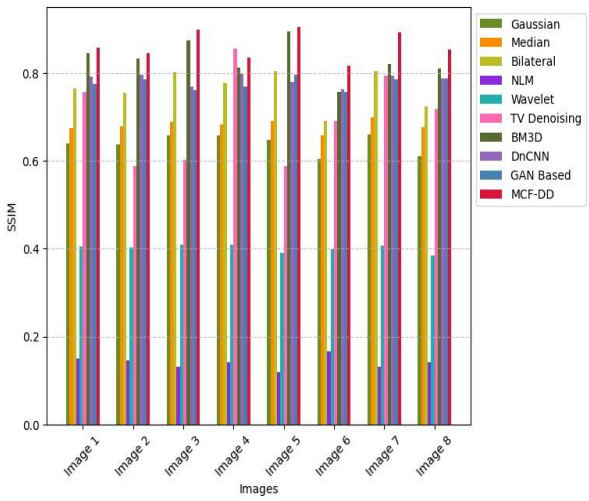


Fig. 6. SSIM comparison of denoising approaches with the proposed method, MCF-DD [Chest X-Ray dataset].

TABLE II. COMPARISON OF DENOISING APPROACHES WITH THE PROPOSED METHOD, MCF-DD [CHEST RAY DATASET].

Denoised Method	Evaluation	Image I	Image II	Image III	Image IV	Image V	Image VI	Image VI	Image V111
	PSNR	27.56	27.10	27.81	26.49	27.17	28.08	27.30	27.38
Noisy Image	SSIM	.4843	.4688	.4786	.4175	.4062	.5747	.4430	.4983
	PSNR	30.30	30.12	30.91	30.21	30.80	29.44	30.69	29.93
Gaussian [13]	SSIM	.6399	.6380	.6585	.6577	.6487	.6057	.6608	.6118
	PSNR	30.72	30.59	31.16	30.36	31.20	30.29	30.98	30.76
Median [14]	SSIM	.6747	.6781	.6893	.6833	.6915	.6589	.6986	.6779
	PSNR	32.70	32.92	33.82	33.11	34.05	31.81	33.67	32.61
Bilateral [15]	SSIM	.7648	.7554	.8030	.7772	.8043	.6914	.8049	.7245
	PSNR	20.39	20.36	20.27	20.15	20.15	20.65	20.17	20.54
NLM [16]	SSIM	.1506	.1454	.1304	.1417	.1178	.1657	.1311	.1421
	PSNR	26.87	26.83	27.05	26.75	26.96	26.71	26.91	26.84
Wavelet [20]	SSIM	.4060	.4023	.4090	.4087	.3909	.3993	.4063	.3853
	PSNR	32.46	29.76	30.23	34.54	30.15	31.04	33.29	32.08
TV Denoising [50]	SSIM	.7570	.5874	.6032	.8564	.5892	.6906	.7944	.7179
	PSNR	34.20	32.68	35.30	33.25	34.84	32.79	34.00	34.01
BM3D [51]	SSIM	.8456	.84343	.8742	.8123	.8943	.7582	.8202	.8099
	PSNR	30.80	31.00	30.25	30.82	31.02	31.08	30.31	31.06
DnCNN [30]	SSIM	.7930	.7958	.7705	.7990	.7800	.7633	.7944	.7874
	PSNR	29.57	29.59	29.61	29.46	29.56	29.46	329.02	29.54
GAN based [31]	SSIM	.7750	.7870	.7610	.7694	.7964	.7583	.7871	.7873
	PSNR	34.96	34.63	36.17	34.23	36.57	33.69	35.70	34.89
MCF-DD	SSIM	.8584	.8452	.8996	.8362	.9062	.8175	.8935	.8536

To understand the better performance of the proposed MCF-DD, the visual results obtained by processing a noisy image using both MCF-DD and other methods are shown in Fig. 7.

Even tough Chest X-Ray Datasets are widely used for medical image analysis, the resultant JPEG compression artifacts may cause unwanted noise that may not be present in the actual image. To perform a more realistic assessment of the performance of the proposed method, uncompressed DICOM images were included from the Medical Imaging Data Resource Center (MIDRC) – RSNA International COVID Radiology Database (RICORD) Release 1c (MIDRC-RICORD-1C) [9]. This dataset, curated by MIDRC and RSNA, consists of high-resolution clinically acquired CXR images stored in DICOM format, preserving the original image quality without lossy compression artifacts. The inclusion of DICOM images allows for a more reliable evaluation of real-world noise characteristics that are typically encountered in medical imaging. In Table III, the comparison of the proposed method with the other methods using images from the (MIDRC) - RSNA International COVID Radiology Database (RICORD) is given.

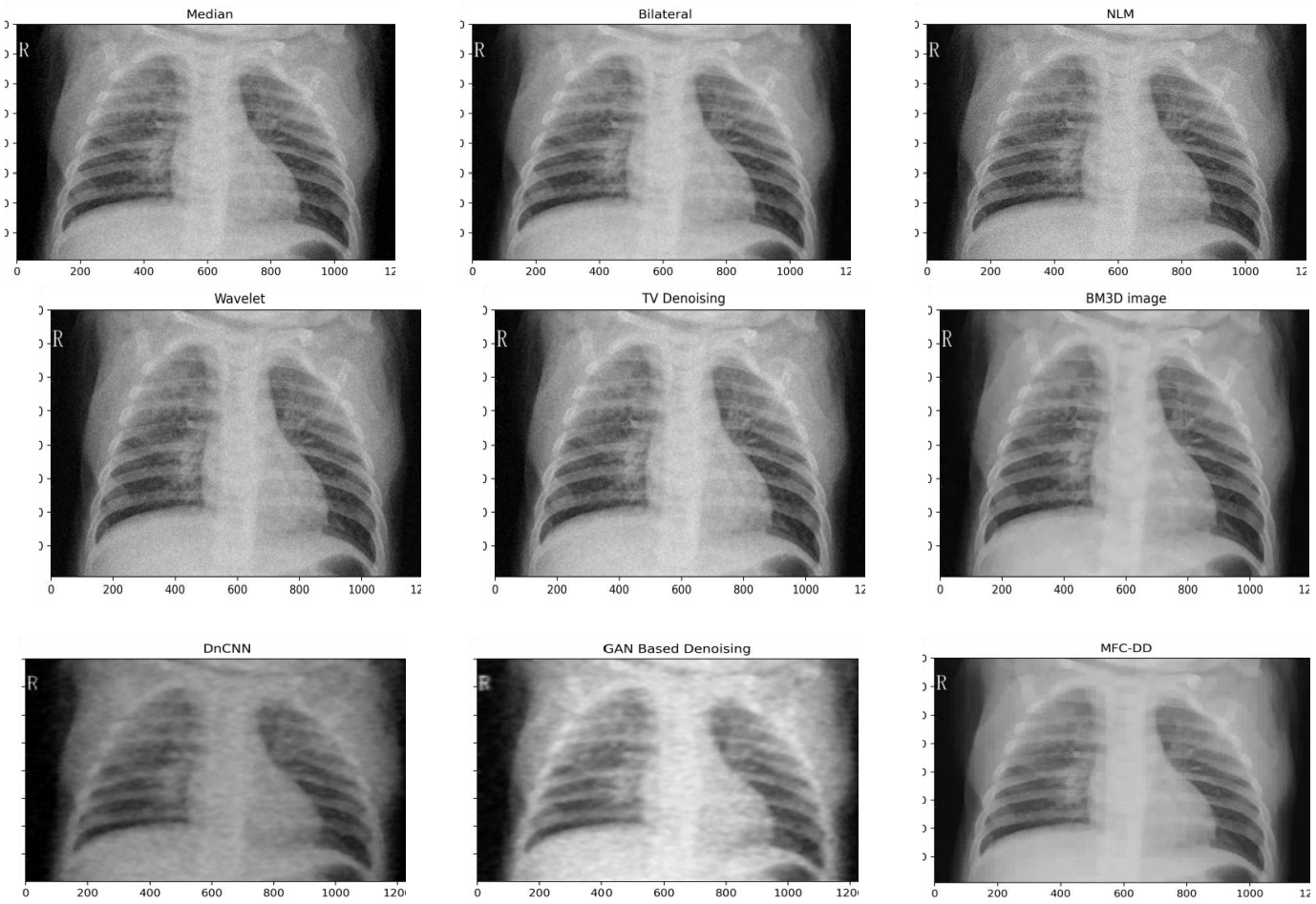


Fig. 7. Visual comparison of denoising methods applied to chest x-ray images [Chest X- Ray Dataset].

TABLE III. COMPARISON OF DENOISING APPROACHES WITH THE PROPOSED METHOD, MCF-DD [(MIDRC) -RSNA INTERNATIONAL 382 COVID RADIOLOGY DATABASE].

Denoised method	Evaluation criteria	Image I	Image II	Image III	Image IV	Image V	Image VI	Image VII	Image VIII
Noisy Image	PSNR	28.21	28.58	27.01	27.93	27.49	25.41	27.17	25.93
	SSIM	.5087	.4704	.5191	.6688	.4714	.2701	.4346	.2824
Gaussian [13]	PSNR	30.07	30.73	29.89	28.33	30.14	36.14	30.50	30.28
	SSIM	.5815	.6327	6390	.4954	.6055	.7685	.6173	.6154
Median [14]	PSNR	31.31	31.33	29.89	28.41	31.11	33.56	31.15	31.07
	SSIM	.6483	.6768	.6259	.4646	.6456	.8009	.6502	.6766
Bilateral [15]	PSNR	32.59	33.65	31.46	29.82	32.91	36.87	33.03	33.98
	SSIM	.7002	.7887	.6959	.5599	.7277	.8097	.7364	.7984
NLM [16]	PSNR	20.48	20.08	20.12	20.64	20.25	30.09	20.14	20.03
	SSIM	.1211	.1153	.1752	.2123	.1298	.2099	.1255	.1972
Wavelet [20]	PSNR	26.85	26.85	26.44	26.15	26.77	31.17	26.73	26.84
	SSIM	.3550	.3755	.4224	.3956	.3763	.6567	.3743	.3507
TV Denoising [50]	PSNR	33.80	33.35	32.31	29.40	29.70	33.45	32.68	29.98
	SSIM	.7790	.7744	.7365	.5375	.5569	.8709	.7313	.550
BM3D [51]	PSNR	33.03	35.90	30.50	29.73	33.06	31.73	34.67	34.38
	SSIM	.8398	.8561	.7161	.5945	.8007	.7589	.8872	.8672
DnCNN [30]	PSNR	30.20	29.67	28.40	29.22	27.71	29.51	30.10	28.11
	SSIM	.7493	.7675	.7343	.7862	.7235	.7634	.7474	.7913
GAN Based [31]	PSNR	27.68	27.40	26.79	27.15	26.86	25.84	27.19	27.02
	SSIM	.7387	.7621	.7340	.7733	.7184	.6984	.7577	.7575
MCF-DD	PSNR	36.34	37.24	32.51	31.01	35.17	38.92	35.51	40.04
	SSIM	.8621	.9089	.7365	.6756	.8267	.9119	.8361	.9476

Fig. 8 and Fig. 9 represent the comparison of the PSNR and SSIM values of the proposed method with the other existing methods for the experiments carried out on DICOM images. From the analysis, we can find that the performance of the

proposed method is better than that of the existing methods. MCF-DD method achieves the highest PSNR and SSIM values across different images. DICOM images are free of compression artifacts. The noise presenting DICOM images represents the original noise characteristics. Conventional denoising techniques have poor performance on DICOM images, while the proposed method adapts well to real X-ray noise patterns, rather than just compression noise.

The proposed MCF-DD (Multi Scale Curvelet based Filtering with Directional Denoising) method demonstrates improved performance for the denoising of chest X-Ray images by combining Curvelet Transform with multiple spatial frequency filters like Gabor, Morlet, and Anisotropic Gaussian, and using a Gaussian Mixture Model (GMM) for adaptive thresholding. The use of Curvelet Transform allows better handling of edges and directional features, which results in multi-scale feature preservation. Also, the combination of different filters helps to capture a wide range of textural and frequency information, contributing to effective noise suppression. The adaptive threshold selection avoids rigid assumptions about the noise model and adapts to local noise characteristics. Table IV presents a comparative analysis of the proposed multi-scale curvelet filtering with the directional denoising technique against existing methods. The results indicate that the proposed approach achieves the highest PSNR and SSIM, demonstrating superior denoising performance and preservation of structural fidelity compared to the state-of-the-art techniques.

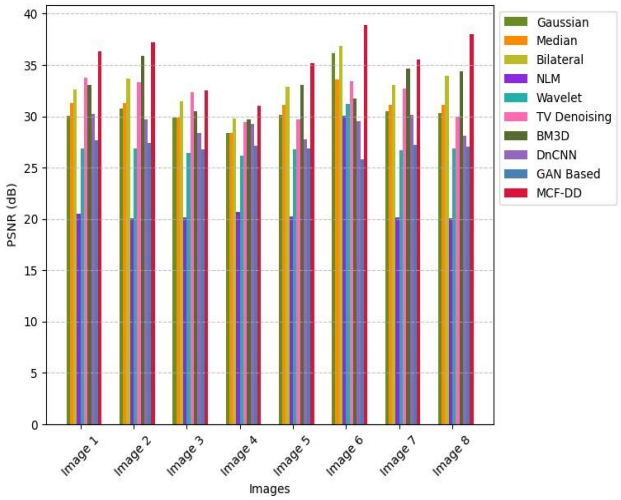


Fig. 8. PSNR comparison of denoising approaches with the proposed method, MCF-DD [(MIDRC) - RSNA International 382 COVID radiology database].

TABLE IV. COMPARISON OF DIFFERENT METHODS ON THE KAGGLE CHEST X-RAY DATASET

Study	Method used	PSNR	SSIM
CNN-based Pneumonia Detection [52]	Convolutional Neural Network (CNNs)	34.12	0.8734
Vision Transformers (ViT) [53]	Vision Transformer (ViT)	35.78	0.8941
Deep Learning Ensemble [54]	Ensemble of GoogLeNet, ResNet-18, DenseNet-121	34.89	0.8812
Hybrid Inception ResNet [55]	Fine-Tuned Inception-ResNet	35.95	0.8994
Proposed method	Multi-Scale Curvelet Filtering with Directional Denoising	**36.57**	**0.9062**

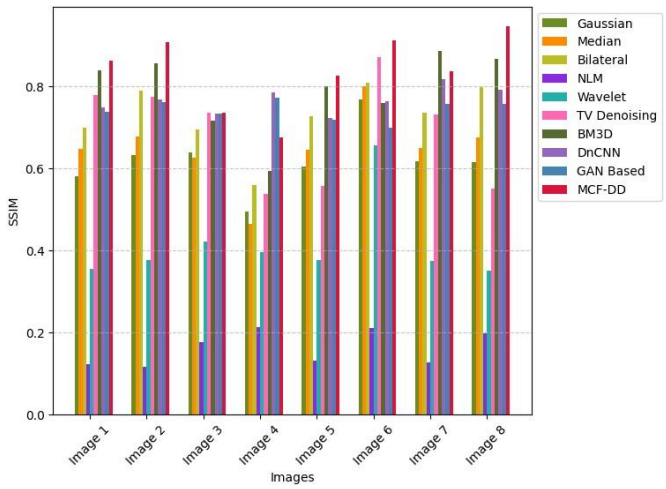


Fig. 9. SSIM comparison of denoising approaches with the proposed method, MCF-DD [(MIDRC) - RSNA International 382 COVID Radiology Database].

The drawbacks of the proposed method include sensitivity to parameter settings and increased computational complexity due to multiple filtering stages. Additionally, this method does not model inter scale dependencies, which may affect performance in highly textured regions. Another limitation of this study is the absence of a radiologist's evaluation to assess whether the denoised images lead to improved diagnostic interpretation. Future work will incorporate domain expert feedback using structured assessment methods, as employed in prior studies [56].

Although PSNR and SSIM are widely adopted for quantitative evaluation of image, they primarily reflect pixel level fidelity and structural similarity. Diagnostic evaluation of the proposed denoising framework has been reported in a previously published study [49], where MCF-DD processed chest X ray images were assessed using standard classification metrics. Accordingly, the present work employs PSNR and SSIM to analyze denoising effectiveness at the signal level.

C. Computational Efficiency Analysis

The runtime comparison presented in Table V indicates that simple local filtering methods such as Gaussian and median filtering have minimal computational cost but provide limited denoising capability. In contrast, non-local and optimization-based approaches, including BM3D and total variation denoising, exhibit substantially higher runtime due to exhaustive block matching and iterative optimization procedures. Learning-based methods reduce this overhead

through feed forward inference but, CNN and GAN-based denoisers still require higher processing time than the proposed framework. With an average inference time of 0.101 ± 0.013 s per image, the proposed MCF-DD method achieves a favorable balance between denoising effectiveness and computational efficiency, remaining significantly faster than BM3D and more efficient than CNN and GAN-based approaches, thereby demonstrating its suitability for large-scale chest X-ray preprocessing.

TABLE V. RUNTIME COMPARISON OF DENOISING METHODS

Method	Runtime (s/image)
Gaussian Filter	0.008 ± 0.009
Median Filter	0.008 ± 0.002
Bilateral Filter	0.075 ± 0.025
Wavelet Denoising	0.083 ± 0.018
Total Variation (TV)	0.760 ± 0.404
BM3D	5.76 ± 0.77
CNN	0.181 ± 0.087
GAN	0.198 ± 0.099
Proposed Method (MCF-DD)	**0.101 ± 0.013**

D. Contribution Analysis of MCF-DD Framework Components.

To understand the contribution of each processing component within the proposed MCF-DD framework, a comparative study was performed by progressively incorporating the framework components. The study began with the Curvelet Transform, which serves as the core multi-scale representation mechanism for noise suppression. Subsequently, the Gabor filter, Morlet filter, and anisotropic Gaussian filter were incrementally integrated into the framework. The average PSNR and SSIM values obtained from the evaluation images are summarized in Table VI.

TABLE VI. MCF-DD FRAMEWORK COMPONENT ANALYSIS

Configuration	PSNR (dB)	SSIM
Curvelet Transform	33.84 ± 0.42	0.861 ± 0.011
Curvelet Transform and Gabor Filter	34.71 ± 0.38	0.882 ± 0.010
Curvelet Transform and Gabor Filter and Morlet Filter	35.48 ± 0.34	0.894 ± 0.008
Curvelet Transform and Gabor Filter and Morlet Filter and Anisotropic Gaussian Filter (MCF-DD)	36.57 ± 0.21	0.906 ± 0.004

The Curvelet Transform alone achieved a PSNR of 33.84 dB and an SSIM of 0.861, demonstrating its effectiveness in suppressing noise while preserving major anatomical structures. When the Gabor filter was incorporated, the PSNR increased to 34.71 Db, and the SSIM improved to 0.882. This improvement can be attributed to the ability of the Gabor filter to capture directional texture information and enhance edge representation. The addition of the Morlet filter further improved performance, resulting in a PSNR of 35.48 dB and an SSIM of 0.894. The Morlet filter enhances multi-scale frequency representation and improves the preservation of subtle anatomical patterns commonly present in chest radiographs. Finally, the integration of the anisotropic Gaussian filter produced the complete MCF-DD framework. This stage effectively suppressed residual noise while preserving edge continuity and structural integrity. Consequently, the proposed MCF-DD framework achieved the highest performance, with a PSNR of 36.57 dB and an SSIM of 0.906.

V. CONCLUSION AND FUTURE SCOPE

In this study, a novel denoising method is introduced, Multi-Scale Curvelet Filtering with Directional Denoising (MCF-DD), along with a patch classification method. This method mainly focuses on the medical chest X-ray images, in which maintenance of structural information of both fine and larger image features is as important as denoising the Poisson and Gaussian noise. A key strength of the method is its dynamic noise classification strategy, powered by a Gaussian Mixture Model (GMM) applied on the variance-to-mean ratio of image patches, allowing adaptive and noise-specific processing. The experimental evaluation demonstrated that MCF-DD outperforms several traditional and state-of-the-art denoising techniques in terms of both quantitative metrics, such as PSNR and SSIM, and qualitative visual quality of the denoised images. The performance has been evaluated on compressed (JPEG) and uncompressed (DICOM) images from diverse datasets, including Kaggle CXR and MIDRC-RICORD.

The adaptability and robustness of the proposed MCF-DD method can be improved by integrating data-driven learning components, such as deep learning-based denoising priors. In addition, the parameter selection process can be automated using optimization algorithms or reinforcement learning frameworks to improve generalization. A critical extension is the integration of the MCF-DD pipeline within explainable AI frameworks to support clinical interpretability.

In summary, MCF-DD presents a flexible, adaptive, and effective solution for image denoising, particularly suited for challenging conditions such as mixed noise scenarios in medical imaging. The promising results lay the foundation for further research and real-world clinical applications.

ACKNOWLEDGMENT

The authors would like to thank the creators of the publicly available chest X-ray datasets used in this study for making the data accessible for research purposes

REFERENCES

[1] M. Diwakar et al., “Low-dose COVID-19 CT image denoising using CNN and its method noise thresholding,” Curr. Med. Imaging Rev., vol. 19, no. 2, pp. 182–193, 2023.

[2] H. Koyuncu and R. Ceylan, “Elimination of white Gaussian noise in arterial phase CT images to bring adrenal tumours into the forefront,” Comput. Med. Imaging Graph., vol. 65, pp. 46–57, 2018.

[3] S. Guo, G. Wang, L. Han, X. Song, and W. Yang, “COVID-19 CT image denoising algorithm based on adaptive threshold and optimized weighted median filter,” Biomed. Signal Process. Control, vol. 75, p. 103552, 2022.

[4] M. Klug et al., "Denoised Ultra-Low-Dose Chest CT to Assess Pneumonia in Individuals Who Are Immunocompromised," Radiol. Cardiothorac. Imaging, vol. 7, no. 2, p. e240189, 2025.

[5] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," Biomed. Signal Process. Control, vol. 61, p. 102036, 2020.

[6] Rashid, N.; Berriri, K.; Albekairi, M.; Kaaniche, K.; Ben Atitallah, A.; Khan, M.A.; El-Hamrawy, O.I. New Real-Time Impulse Noise Removal Method Applied to Chest X-ray Images. Diagnostics 2022, 12, 2738. https://doi.org/10.3390/diagnostics12112738.

[7] J. Mohan, V. Krishnaveni, and Y. Guo, "A survey on the magnetic resonance image denoising methods," Biomed. Signal Process. Control, vol. 9, pp. 56–69, 2014.

[8] D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Cell, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/J.CELL.2018.02.010.

[9] E. Tsai et al., "Data from Medical Imaging Data Resource Center (MIDRC) – RSNA International COVID Radiology Database (RICORD) Release 1c – Chest x-ray, Covid+ (MIDRC-RICORD-1C)." The Cancer Imaging Archive, 2021, doi: 10.7937/91ah-v663.

[10] M. Diwakar and M. Kumar, "A review on CT image noise and its denoising," Biomed. Signal Process. Control, vol. 42, pp. 73–88, 2018.

[11] D. K. Priya, B. B. Sam, S. Lavanya, and A. P. Sajin, "A survey on medical image denoising using optimisation technique and classification," in 2017 international conference on information communication and embedded systems (ICICES), 2017, pp. 1–6, doi: 10.1109/ICICES.2017.8070729.

[12] L. Fan, F. Zhang, H. Fan, and C. Zhang, "Brief review of image denoising techniques," Vis. Comput. Ind. Biomed. Art, vol. 2, no. 1, p. 7, 2019.

[13] P. Chen, H. Qian, and M. Zhu, "Fast Gaussian particle filtering algorithm [J]," J. Huazhong Univ. Sci. \& Technol., 2008.

[14] T. Huang, G. Yang, and G. Tang, "A fast two-dimensional median filtering algorithm," IEEE Trans. Acoust., vol. 27, no. 1, pp. 13–18, 1979.

[15] D. Bhonsle, V. Chandra, and G. R. Sinha, "Medical image denoising using bilateral filter," Int. J. Image, Graph. Signal Process., vol. 4, no. 6, p. 36, 2012.

[16] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 2005, vol. 2, pp. 60–65, [Online]. Available: 10.1109/CVPR.2005.38.

[17] F. Jin, P. Fieguth, L. Winger, and E. Jernigan, "Adaptive Wiener filtering of noisy images and image sequences," in Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), 2003, vol. 3, pp. III--349.

[18] T. Barbu, "Robust anisotropic diffusion scheme for image noise removal," Procedia Comput. Sci., vol. 35, pp. 522–530, 2014.

[19] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," IEEE Trans. image Process., vol. 9, no. 9, pp. 1532–1546, 2000.

[20] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: asymptopia?," J. R. Stat. Soc. Ser. B, vol. 57, no. 2, pp. 301–337, 1995.

[21] S. Hashemi and S. Beheshti, "Adaptive image denoising by rigorous Bayesshrink thresholding," in 2011 IEEE Statistical Signal Processing Workshop (SSP), Jun. 2011, pp. 713–716, doi: 10.1109/SSP.2011.5967802.

[22] M. S. Moad, N. Zermi, A. Khaldi, and M. R. Kafi, "Stationary wavelet-based image watermarking for E-healthcare applications," Cybern. Syst., vol. 56, no. 2, pp. 81–96, 2025.

[23] F. E. Ali, I. M. El-Dokany, A. A. Saad, and F. E. Abd El-Samie, "A curvelet transform approach for the fusion of MR and CT images," J. Mod. Opt., vol. 57, no. 4, pp. 273–286, 2010.

[24] J. K. Appati, B. Ziamah, H. A. Akrofi, and A. A. Dodoo, "SARS detection in chest CT scan images using the bootstrapped ViT-B/16 model," Iran J. Comput. Sci., pp. 1–15, 2025.

[25] Z. Zhang, J. Cui, X. Luo, and Q. You, "Statistical correlative model in the multimodal fusion of brain images," Int. J. Imaging Syst. Technol., vol. 30, no. 4, pp. 1066–1079, 2020.

[26] K. and K. F. and C. T. and K. A. and C. L. Zotin Alexanderand Simonov, "Techniques for Medical Images Processing Using Shearlet Transform and Color Coding," in Computer Vision in Control Systems-4: Real Life Applications, L. C. Favorskaya Margarita N.and Jain, Ed. Cham: Springer International Publishing, 2018, pp. 223–259.

[27] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," IEEE Trans. image Process., vol. 26, no. 7, pp. 3142–3155, 2017.

[28] J. Lehtinen, "Noise2Noise: Learning Image Restoration without Clean Data," arXiv Prepr. arXiv1803.04189, 2018.

[29] M. Nishio et al., "Convolutional auto-encoder for image denoising of ultra-low-dose CT," Heliyon, vol. 3, no. 8, 2017.

[30] Murali Vineeth and P. V and Sudeep, "Image Denoising Using DnCNN: An Exploration Study," in Advances in Communication Systems and Networks, Jun. 2020, pp. 847–859, doi: https://doi.org/10.1007/978-981-15-3992-3_72.

[31] H.-J. Kim and D. Lee, "Image denoising with conditional generative adversarial networks (CGAN) in low dose chest images," Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip., vol. 954, p. 161914, 2020.

[32] M. Sandhu, S. Kushwaha, and T. Arora, "A Comprehensive Review of GAN-Based Denoising Models for Low-Dose Computed Tomography Images," Int. J. Image Graph., p. 2550030, 2023.

[33] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," IEEE Trans. Image Process., vol. 16, no. 8, pp. 2080–2095, 2007, doi: 10.1109/TIP.2007.901238.

[34] F. Deeba, S. Kun, F. A. Dharejo, and Y. Zhou, "Wavelet-based enhanced medical image super resolution," IEEE Access, vol. 8, pp. 37035–37044, 2020.

[35] N. Mohana Suganthi and M. Arun, "Diabetic retinopathy grading using curvelet CNN with optimized SSO activations and wavelet-based image enhancement," Ain Shams Eng. J., vol. 16, no. 1, p. 103239, 2025, doi: 10.1016/j.asej.2024.103239.

[36] M. Thelwall, "The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach," J. Informetr., vol. 10, no. 1, pp. 110–123, Feb. 2016, doi: 10.1016/j.joi.2015.12.001.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. image Process., vol. 13, no. 4, pp. 600–612, 2004.

[38] M. Qutbi, "On the Convergence of Maximum Likelihood Expectation--maximization Algorithm for Iterative Tomographic Reconstruction and the Role of Spatial Frequency, Location, Pixel Intensity, Contrast, and Noise," J. Med. Phys., vol. 50, no. 2, pp. 348–358, 2025.

[39] F. J. Anscombe, "The Transformation of Poisson, Binomial and Negative-Binomial Data," Biometrika, vol. 35, no. 3–4, pp. 246–254, Dec. 1948, doi: 10.1093/biomet/35.3-4.246.

[40] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," Vision Res., vol. 20, no. 10, pp. 847–856, 1980.

[41] J. Zou, C.-C. Liu, Y. Zhang, and G.-F. Lu, "Object recognition using Gabor co-occurrence similarity," Pattern Recognit., vol. 46, no. 1, pp. 434–448, 2013.

[42] J. Morlet, G. Arens, E. Fourgeau, and D. Glard, "Wave propagation and sampling theory—Part I: Complex signal and scattering in multilayered media," Geophysics, vol. 47, no. 2, pp. 203–221, 1982.

[43] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 7, pp. 629–639, 1990.

[44] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," Biometrika, vol. 81, no. 3, pp. 425–455, 1994.

[45] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities," Commun. Pure Appl. Math., vol. 57, no. 2, pp. 0219–0266, Feb. 2004, doi: 10.1002/cpa.10116.

[46] E. Candès, L. Demanet, D. Donoho, and L. Ying, "Fast Discrete Curvelet Transforms," Multiscale Model. \& Simul., vol. 5, no. 3, pp. 861–899, 2006, doi: 10.1137/05064182X.

[47] H. Wei and W. Zheng, "Image denoising based on improved gaussian mixture model," Sci. Program., vol. 2021, no. 1, p. 7982645, 2021.

[48] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[49] N. Sebastian and B. Ankayarkanni, "Enhanced ResNet-50 with Multi-Feature Fusion for Robust Detection of Pneumonia in Chest X-Ray Images," Diagnostics, vol. 15, no. 16, p. 2041, 2025.

[50] B. Li and D. S. Que, "Medical Images Denoising Based on Total Variation Algorithm," Procedia Environ. Sci., vol. 8, pp. 227–234, Jan. 2011, doi: 10.1016/j.proenv.2011.10.037.

[51] I. Djurović, "BM3D filter in salt-and-pepper noise removal," EURASIP J. Image Video Process., vol. 2016, pp. 1–11, 2016.

[52] M. S. Al Reshan et al., "Detection of pneumonia from chest X-ray images utilizing mobilenet model," in Healthcare, 2023, vol. 11, no. 11, p. 1561.

[53] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," Sci. Rep., vol. 14, no. 1, p. 2487, 2024.

[54] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," PLoS One, vol. 16, no. 9, p. e0256630, 2021.

[55] M. Neshat, M. Ahmed, H. Askari, M. Thilakaratne, and S. Mirjalili, "Hybrid inception architecture with residual connection: fine-tuned Inception-ResNet deep learning model for lung inflammation diagnosis from chest radiographs," Procedia Comput. Sci., vol. 235, pp. 1841–1850, 2024.

[56] [G. Mehdipoor, F. Salmani, and A. Arjmand Shabestari, "Survey of practitioners' competency for diagnosis of acute diseases manifest on chest X-ray," BMC Med. Imaging, vol. 17, pp. 1–6, 2017.

Characterizing Failure Points in Rule-Based Spreadsheet Data Transformation: A Stage-Oriented Taxonomy and Empirical Failure Matrix

Fakhrul Adli Mohd Zaki, Mustafa Man, Mohamad Nor Hassan
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Terengganu, Malaysia

***Abstract*—Rule-based spreadsheet data transformation remains widely used for converting human-centred spreadsheet tables into structured formats for reporting, analytics, and data integration because it is transparent, auditable, and reproducible. However, rule-based approaches often fail when spreadsheets encode structural meaning through visual or layout cues, such as multi-row headers, merged cells, interleaved subtotals, cross-tabulated values, and multiple logical tables within a worksheet. This study characterizes such failures using a stage-oriented taxonomy aligned with a six-stage transformation pipeline: ingestion, region detection, structural interpretation, semantic labeling, canonicalization, and validation. A deterministic baseline pipeline was evaluated against six controlled spreadsheet cases representing common semi-structured layouts. The resulting empirical failure matrix identified 15 severity-2 failures across the controlled cases. The most critical failures were concentrated in merged cell and hierarchy handling, multi-region and boundary interpretation, role mislabeling, wrong granularity, and silent semantic errors. These findings show that many transformation failures are not isolated rule defects but arise from missing pipeline capabilities, particularly explicit hierarchy representation, block-level segmentation, granularity control, and semantic invariant checking. The study contributes a reproducible failure-analysis framework that supports stage-localized diagnosis, comparison of rule-based spreadsheet transformation systems, and targeted improvement of transformation pipelines.**

Keywords*—Rule-based transformation; spreadsheet table understanding; data wrangling; failure taxonomy; failure matrix; canonicalization

I. Introduction

This study examines a necessary first step toward improving rule-based spreadsheet transformation: identifying how and where such transformers fail. Rather than proposing a new model or algorithm, the study provides a structured failure analysis. Specifically, it proposes 1) a failure taxonomy that maps failures to stages in a common rule-based transformation process and 2) a failure matrix derived from controlled test cases (S1-S6) that represent common spreadsheet layout patterns. This analysis supports more comparable failure reporting, improves rule-writing practices, and helps future designers create more robust validation and debugging methods.

The study makes three main contributions. First, it introduces a stage-oriented taxonomy that maps failures to components of a typical rule-based pipeline, including interpretation of the input spreadsheet, rule selection, rule execution, and output validation. Instead of presenting only a long list of error symptoms, the taxonomy supports failure diagnosis, system comparison, and structured reporting of transformation problems. Second, it presents an empirical failure matrix with measured severity scores. The matrix represents each identified failure type against six controlled spreadsheet test-case patterns (S1-S6), documenting which failures are triggered by particular structural or semantic issues and which case-code pairs produce the highest-severity failures. This representation is easier to interpret than isolated error lists because it highlights recurring and under-reported failure categories while grounding each failure in a concrete example. Third, it derives practical implications for improving rule writing and pipeline design. The results suggest the need for additional checkpoints to detect silent errors, more structure-aware interpretation of spreadsheet semantics, and explicit handling of multiline headers, grouped headers, merged cells, and subtotal rows.

A. Practical Motivation and Scope

Spreadsheet tables are designed for human users and often rely on a variety of non-schema-based cues for layout and data semantics. Such cues include consistent indentation, spacing, cell merging, bold text, borders, and footnotes. Automating the interpretation of spreadsheet tables using only the content of cells and cells' properties proves challenging because they lack formal semantics defined externally. Many researchers attempt to overcome these limitations by inferring table semantics from visual cues and their context. [6], [7]

In practice, rule-based scripts are commonly utilized to extract recurring table formats, such as monthly financial reports, departmental budgets, customer purchase orders, and survey responses. The strengths of rule-based conversion scripts are their transparency, auditability, maintainability with version control, reusability across different reports, and suitability for batch processing. One approach to productionizing rule-based conversion is through rule-based platforms, which apply a set of predefined transformation rules to the structured table data. These platforms ensure consistent and repeatable processing, but first, the rules for a particular table must be correctly defined to reflect its structure. [7]

However, many rule-based conversion approaches implicitly assume that tables have a uniform structure throughout their scope, which is a fragile assumption. First, it generally requires that a table consist of a single continuous

block. Second, it may require that the header depth be constant throughout. Third, if the cells in the table were merged to make reading more convenient for users, then the merging should be consistently applied and handled in the script. Fourth, it generally requires that subtotals and totals are clearly marked (perhaps using bold font or a common text string such as "Total") so they can be identified easily by filtering rows based on that specific visual cue. When tables are not consistent in these respects, rule-based conversion either will fail to process the spreadsheet table altogether, resulting in no output, or it will correctly process the input rows but generate incorrect table outputs, even though the outputs may be syntactically valid. These are insidious errors because the bad output might be used in downstream data analysis without producing any visible errors. Work in previous rule-based techniques shows that minor changes to the layout (e.g., adding a bold font) are sufficient to change the semantic meaning of the table, leading to a functionally different table from the one that the user is looking at in the spreadsheet. [6], [7]

This study is about failure modes of rule-based spreadsheet data extraction. Instead of introducing a new extraction technology or a novel machine learning technique for such a purpose, this study is an attempt to provide structure to the discussion around different failure modes of rule-based systems in spreadsheet data extraction. Designers of rule pipelines for such purposes should use this characterization to understand where failures come from, how they propagate through the pipeline, and how they might be fixed. This study first provides a stage-specific taxonomy of failures in rule-based extraction systems and then presents an empirically measured failure matrix based on a set of carefully chosen case spreadsheets exhibiting common layout types. This allows for a concise summary of the coverage of empirical experiments and recurring weaknesses that have been identified in such systems. [5]

B. Study Organization

The remainder of this study is divided into nine sections. Section II provides a survey of prior research, focusing on interactive data wrangling, programming by example, spreadsheet table understanding, and rule-based transformations. Section III details the formal setup for transformations and introduces the concept of a transformation failure point. Section IV outlines the architecture of the baseline data transformation pipeline, defines the associated set of basic rules, describes the controlled cases created for the experiment, and explains the metrics and procedure used for evaluating outcomes. Section V elaborates on the comprehensive taxonomy of failure types encountered during transformation. Section VI presents the constructed failure matrix, along with calculated severity scores, and discusses the observed trends and prominent types of failures. Section VII analyzes the identified failure patterns and proposes strategies for more robust rule creation. Section VIII addresses potential limitations and threats to the study's validity. Finally, Section IX summarizes the key findings and suggests avenues for future research.

II. Related Work

This section covers previous research on the manipulation and preprocessing of data stored in spreadsheets. The literature review is divided into four categories: data wrangling and interactive transformation, programming-by-example, table understanding and rule-based transformations, and data preparation validation/error detection. The categories are related due to the inherent nature of spreadsheet transformations; a transformation is often not simply changing data format but also inferring spreadsheet layout, interpreting tabular structure, maintaining semantic relations, and creating an output that is accurate and useful. The review explores how different techniques empower users to clean, transform, extract, and validate data in spreadsheets and highlights challenges that remain with current methods.

A. Data Wrangling and Interactive Transformation

Research findings demonstrate that data wrangling is a repeated activity during which users perform certain transformations and then observe the outputs. This interactive cycle involves letting users visualize transformation outputs and allowing them to reverse actions if outputs are invalid. Petricek et al. built the notion of AI assistants for data wrangling upon this idea; users supply constraints to the assistant to obtain a system-generated transformation satisfying constraints and minimizing wasted effort through trial and error to carry out practical cleaning and reshaping operations [1]. In practice, Wrangler is perhaps the best example of this kind of tool, being another mixed-initiative approach, which blends direct manipulation of data with an automatically inferred sequence of transformation steps, allowing analysts to incrementally craft the transformations for operations like splitting a field, removing noise, and transforming a column’s appearance, all while keeping the transformation transparent [2]. Despite the gains in usability these tools provide, they encounter challenges with implicit structural cues such as the use of merged headers for defining sections, separate groups of cells to define regions or sections of data, and lines/ separators to make report layouts clearer, which reinforces the need for more robust structure treatment and automated validation in ETL (Extract, Transform, Load) systems [5].

B. Programming by Example in Spreadsheets

Learning transformations from examples, known as programming-by-example methods, greatly reduces the burden of program specification. They perform best at cell-level and string-level operations common in typical spreadsheet manipulation. In [3], Gulwani et al. provide a survey on how to do data manipulations in spreadsheets using examples. The study demonstrates the effectiveness of program synthesis with examples to learn both simple string and complex cell transformation tasks that would otherwise be inefficient to write with manually defined formulas or scripts. However, the vast majority of spreadsheet data manipulations deal not just with value transformation but with transformation of spreadsheet structure itself, like reorganizing data from a crosstab format into record-based rows, maintaining hierarchical headers spanning multiple rows, or separating

subtotals from data points. To support such transformations, one must often deduce implicit relationships among values and recover latent table structure and semantic roles, which are underspecified by only a limited number of input-output examples and depend on a wide range of cues, such as layout, visual appearance, and semantics associated with values or context [5], [6].

C. Spreadsheet Table Understanding, Extraction, and Rule-Based Transformation

The research into spreadsheet table understanding focuses on identifying specific regions within a spreadsheet table, recognizing its headers, and assigning roles to each cell in order to represent the grid as a meaningful logical structure that can be transformed. Building upon this, rule-based transformations apply rules that explicitly encode human interpretation steps in analysis and processing. Shigarov and Mikhailov present a rule-based technique coupled with a table object model that creates two conceptual layers: one to store physical cell information and another to contain logical semantic items. This layering facilitates the transformation of varied tables into a standardized relational format [4]. Numerous rule-engine implementations exist that demonstrate how to use production rules to analyze table formatting, content in natural language, and spatial structure to derive meaningful relations between labels and their associated entries, then use such derived relationships to guide the generation of standard tables [6]. Recently, TabbyXL commercialized this vision by incorporating logical items, such as entries, labels, and categories, into a table object model, allowing user-written rules to be executed either by an engine that conforms to the JSR-94 specification (e.g., Drools) or by transforming a tailored rule language known as CRL into actual code for execution as Java programs [7]. While these works focus primarily on parsing and analysis, other related framework work points out the full requirements for a comprehensive extraction system. A robust extraction system must also handle data cleanup, information provenance/lineage tracking, and generation of relational and potentially linked data. This suggests the necessity for a uniform workflow that accommodates the diverse array of table structures [5].

D. Validation and Error Detection in Data Preparation

One of the largest problems in data preparation is that we can never confirm the correctness of output; countless variations are acceptable on the surface, each a plausible candidate depending on specific implicit or explicit constraints. In light of this, data preparation solutions generally suggest employing validation checks, error identification, and interactive assistance. Validations might simply involve executing schema type and constraint checks, but could extend to sophisticated domain rules and checks that verify reconciliation and logical consistencies. This challenge plays itself out in the humble spreadsheet, which has implicit rules obvious to a user but unstated: We should not export a row labelled "total" as an observation, column headers should properly identify the contents of each group in a group-and-sort operation, a column of names should not have duplicates. This is addressed in rule engines with explicit constraints to reject or filter records, for example, rejecting records whose first field contains "total" [6]. Another strategy is used in platform-based solutions: the comparison of transformed data against source truth, which checks whether required "Functional Items" have been extracted, whether the relationships extracted reflect source truth, etc. [7]. A problem with this is that it differentiates a failed operation that stops the process from producing output and an operation that, in effect, creates data that looks valid on the surface but is semantically garbage. Failure detection would ideally highlight errors in a repeatable manner, identifying specifically repeatable tests that can be built into a repeatable data transformation [5].

Existing studies have laid important foundations for spreadsheet data wrangling, programming-by-example, table understanding, rule-based interpretation, and validation [1]–[8], [10], [11]. These studies mainly focus on transformation support, table extraction, semantic interpretation, or output generation, rather than systematic failure localization across different stages of a spreadsheet transformation pipeline. This limitation is particularly important for spreadsheet data because spreadsheet meaning is often embedded in structural and visual cues, such as merged headers, section labels, subtotals, and multiple regions of related data [4]–[7], [10], [11]. Interactive wrangling and programming-by-example approaches reduce the burden of specifying spreadsheet transformations, but they provide limited explanation of where structural or semantic failures originate during the transformation process [1]–[3]. Rule-based table understanding approaches offer transparent and auditable mechanisms for extracting and transforming spreadsheet information, but their evaluations are commonly organized around specific interpretation or extraction tasks, such as header mapping, entity relationship extraction, or table-to-relational conversion, rather than a systematic stage-wise failure taxonomy [4]–[8]. Recent machine learning and transformer-based approaches demonstrate the importance of learning internal representations for table structure and semantics, including relational table representation, spatial structure, hierarchical structure, and tabular representation learning [12]–[16]. Nevertheless, rule-based table understanding remains relevant for task-specific transformation scenarios that require deterministic, auditable, reproducible, and domain-specific solutions [4], [6], [7]. Most existing approaches either address transformation for general data sources, focus on spreadsheet extraction, or emphasize validation after transformation. To date, limited attention has been given to stage-oriented failure characterization that systematically maps spreadsheet layout patterns, transformation pipeline stages, failure types, empirical severity scores, and corresponding examples. Therefore, this study addresses the gap in systematically characterizing stage-wise failures and their relationship with spreadsheet layout patterns, pipeline stages, failure types, severity levels, and concrete failure examples.

III. Problem Definition

A worksheet is defined as the grid C of cells, where each cell is an entity with a specific coordinate (r,c), value v, optional formula f, and associated visual features like borders and colorization g. A function T, defined as the rule-based pipeline, converts this worksheet S into a structured dataset D, the schema of which can be varied. Two critical processes in understanding table information are identifying which cells

play which functional role (i.e., metadata, headers, labels, data cells, or derived cells) and determining the table layout. This process precedes and makes possible the transformation step [10], [11].

A transformation pipeline failure point is the stage at which the applied rule set produces one of the following conditions:

- No output is produced when output is expected, for example, when the pipeline fails to detect any table region.
- The output is structurally incorrect, for example, when the detected region boundaries are wrong or when the predicted headers and data cells do not correspond to the intended table structure.
- The output is semantically incorrect, for example, when the extracted field names are wrong, the inferred data structure does not reflect the table structure, or the record granularity does not correspond to the perceived entities in the source table.

We can distinguish between overt pipeline failures (e.g., pipeline crash or empty return) and silent failures (e.g., the pipeline returns syntactically correct but conceptually incorrect results). Silent failures are the more problematic type for the automatic pipeline process.

The pipeline transformation *T* function described can, in practice, be understood as a set of stages that process and refine the intermediate table structure. Earlier stages aim to segment cells belonging to a potential table. Middle stages deduce header depth and structure and demarcate data regions. Finally, later stages arrange cells in canonical records and enforce data validity rules. This staged pipeline view is significant for diagnosing failure sources because failures often lead to subsequent failures: incorrect boundaries lead to incorrect headers, which lead to wrong field names, leading to wrong record keys and incorrect granularity. Accordingly, the failure diagnostic should record the specific stage at which a problem is introduced into the internal table representation, in addition to the stage at which it first becomes visible in the generated output.

IV. Methodology

Following Section III, the transformation *T* is defined as a deterministic sequential operation that takes a worksheet *S* and produces either a dataset D_{out} or an empty output accompanied by an error message. The multistage pipeline isolates where errors arise and when they become visible in the output.

A. Pipeline Stages

The proposed model consists of six stages, labelled A to F. At each stage, the pipeline processes an intermediate representation of the worksheet, produces an output, and records diagnostics. The staged view supports debugging because an early structural issue, such as failure to recognize a table header at Stage B, may later appear as a semantic error in Stage E. Table parsing also benefits from staged decomposition because visual layout cues can convey semantics that cannot be captured by schema interpretation alone [10]. Rule-based table extraction systems also estimate semantic links between cells from spatial, typographic, and textual cues, making staged diagnostics a natural output of such systems [6]. Existing rule engines expose actions for analysis, semantic interpretation, canonical output generation, and evaluation, which makes staged processing useful for both development and reproducible debugging [7], [5].

Stage A: Ingestion and Normalization. Cleans the cells: removes blank spaces at the start and end of a cell text, converts numerical values to a common format when possible, replaces spreadsheet formulas with their results when possible, records merged cells as objects that retain the anchor and span coordinates, and documents merge operations.

Stage B: Region Detection. Scans the worksheet and identifies rectangular regions, then filters and ranks them based on heuristics related to the density of text and contiguity. Output is a ranked list of candidate regions; the best region R, is prioritized for subsequent processing.

Stage C: Structural Interpretation. Estimates the depth k of the header, partitioning the region R into a header and a body, and infers structure patterns such as a cross-tabulated or a multi-header table section or other organizational structures within the table.

Stage D: Semantic Labelling. Identifies a row or column or a cell as a specific kind of label or data, or an annotation, and associates header labels with fields that'll be produced in the output. Uses blank value labelling rules and normalization for strings.

Stage E: Canonicalization. Converts the internal table structure into a flat, relational table structure, flattening hierarchical headers, defining keys for records, and potentially dropping non-observation rows such as subtotals.

Stage F: Validation. Validates the output against schema constraints, datatype constraints, and key constraints, and adds plausibility checks, such as an acceptable range for the number of records. Rejects clearly malformed results and highlights potential problems.

B. Baseline Rule Set

For this study, a representative rule set comprising six deterministic rules, denoted R1-R6, was developed to demonstrate the failure characterization technique. The rules were not designed to optimize table accuracy metrics; instead, they serve as diagnostic heuristics that reveal frequent failures when spreadsheet semantics are communicated through layout patterns. These rules resemble generic heuristic rules commonly found in industrial rule-based pipelines and can be improved after specific failures are identified [6], [8]. The same baseline rules are applied to all controlled cases to avoid confounding the results with rule variations and to ensure that the observed failures are attributable to layout patterns.

R1 (Region Selection): This rule detects and selects the densest rectangular region of non-empty cells to be treated as a candidate table.

R2 (Header Depth Inference): This rule estimates the depth of the table header (k) by heuristically inferring the depth of numerical columns, but it is capped at a maximum value.

R3 (Merged-Cell Propagation): This rule propagates the content of merged cells among the cells that they span.

R4 (Body Extraction): This rule extracts the table body, defined as the region of the spreadsheet directly below the estimated header.

R5 (Header Flattening): This rule "flattens" hierarchical table headers, such as when nested cells are used to represent broader categories. It accomplishes this by concatenating the cell content along each header path, delimited by forward slashes.

R6 (Subtotal Filtering): This rule eliminates rows that are identified as containing subtotal data (e.g., rows containing terms such as "Total" or "Subtotal") based on a simple keyword-matching pattern.

In our empirical study, we run these rules against each case in the controlled experiments and log all failures that arise. These failures are categorized based on our classification system of failure codes. We populate the failure matrix by noting the severity of the highest failure observed for each case-code pair. This matrix thus summarizes the failure pattern for the baseline rules on all of our carefully crafted test cases [7].

C. Controlled Spreadsheet Cases

Six controlled spreadsheet cases (S1-S6) were used to represent common semi-structured layout patterns that challenge rule-based approaches. The cases differ in three ways: 1) each contains a single worksheet, 2) each includes layout artifacts that must be handled, such as merged headers, interleaved subtotals, crosstab layouts, or multiple table regions, and 3) each has a ground-truth specification identifying the primary data region, header depth, fields, and observation formation rules, including the exclusion of subtotal rows (Table I).

TABLE I. CONTROLLED SPREADSHEET CASES

Case	Pattern	Key Characteristic	Expected Output Type
S1	Simple tidy table	Single header row, contiguous region, no merges	Direct row-to-record mapping
S2	Multi-row header	2-3 header rows, hierarchical column labels	Composite field construction from the header hierarchy
S3	Crosstab	Row categories × column categories with values	Reshape to relational (long-form records)
S4	Merged-cell semantics	Merged headers/group labels encode hierarchy	Hierarchy-preserving header interpretation
S5	Interleaved subtotals	Subtotal rows within body, mixed numeric and text	Granularity control (filter non-observations)
S6	Multi-region sheet	Multiple tables plus notes or separators	Region segmentation and correct table selection

The expected output for a single observation record is the corresponding record within a relational database table, stripped of decorative titles, any spacing lines, notes associated with the data, and subtotals or totals.

For tables that span multiple rows within the header region, the concatenated header fields for each data point become the singular header label for the data, and this concatenation follows a specific, consistent ordering (matching R5). Cross-tabulated data are transformed into a long-form configuration in which the reported values each represent a corresponding cell identified by their respective row and column values.

D. Evaluation and Failure Scoring Procedure

Checks were implemented aligned with the three outcome categories from Section III to operationalize failure reporting using taxonomy codes (Section III). For each check, we record a binary pass/fail status and a descriptive diagnostic message that indicates which check was performed, when it became apparent in the pipeline (stage), and which baseline rule identifiers are involved. For a particular table, the failure matrix (Table III) sums the severity score across all failed checks for each taxonomy code. An occurrence is thus characterized by the tuple (table ID, taxonomy code) for which the failure matrix reports the highest severity score.

No-output failures. We identify a no-output failure when the pipeline outputs an empty table (while there is a valid input table). This can occur either if Stage B fails to find at least one candidate exceeding a given threshold or if Stage F rejects all produced candidate regions. A no-output failure likely indicates issues arising during the region discovery stage (Stage B) and enforced by the selection process (Stage F).

Structural incorrectness. We consider the output to be structurally incorrect when the extracted region differs from the specified or intended region. Additionally, we note a structural error when the automatically derived header depth lies outside the range of [0, max header depth] or when the output structure violates basic requirements, such as the absence of essential columns or inconsistencies between headers and bodies. Failures of this nature are predominantly associated with problems in stages B and C.

Semantic incorrectness. We deem the output semantically incorrect when it breaks assumptions about its correct interpretation and meaning, despite possessing a valid structure. This encompasses various issues: null values or composite key incoherence, duplicated keys resulting from improper propagation of label interpretations, omission of row-level summary figures (subtotals), and errors in flattening cross-tabulations, which may misassign values to cell-group combinations. This classification explicitly encompasses subtle semantic problems that do not trigger schema checks but still violate semantic constraints. Semantic errors frequently emerge from issues in stages D and E and become observable during Stage F.

Severity scoring. We quantify the severity of each recorded failure: 0 (no failure), 1 (partial degradation, fixable with a localized rule modification), and 2 (critical failure requiring new capability, e.g., reasoning across multiple regions, explicitly representing a hierarchy, etc.). These scores form the entries in the failure matrix in Section VI.

E. Pseudo Algorithm (Algorithm 1)

Algorithm 1: Stage-wise transformation and failure logging

Input: Worksheet S; baseline rules R1..R6; failure taxonomy FA2-FF1.

Output: Extracted dataset Dout (or empty) and failure log L

Normalize S (clean cell content; represent merged spans).

Detect candidate regions; select R using R1.

If no valid region is selected, log a no-output failure (taxonomy code in Stage B) and return (emptyset, L).

Infer header depth k using R2; partition R into header/body.

Propagate merged header labels using R3; assign roles/labels and construct field labels.

Flatten and canonicalize using R4–R6 (including hierarchical header flattening and subtotal filtering).

Validate output using schema/key/plausibility checks.

For each failed check, append to L a record Failure logging:

For each failed check, append to L a record (case, taxonomy code, outcome type, stage visible, severity, diagnostic), where outcome type is one of {no-output, structural, semantic} and severity is one of {0,1,2}.

Return: Output (Dout, L). The failure matrix (Table III) is populated by taking the maximum severity for each case (case, taxonomy code) pair.

V. Failure Taxonomy

Existing approaches for spreadsheet table understanding or rule-based transformation largely focus on region detection, role extraction, table semantics, or simply generate relational outputs, while this taxonomy aims to localise failure and describe failure propagation during the pipeline. It not only labels the visible symptom in an incorrect output but also helps to determine at which stage of processing the error occurred because failure might originate during region detection or interpretation stage, but become visible later, only during normalisation or validation, such as invalid boundaries show up as error during semantics labelling or remain unseen as silent failure during validation. The taxonomy is therefore a useful addition to the previous work on table understanding, which serves a diagnostic purpose, linking the types of failure, the stages of the pipeline where they originate, the patterns that trigger them in the spreadsheet, and the effects they cause in the output result. This stage-oriented view also provides for a fairer or more reasonable comparison between the rule-based pipelines because failures can be reported in terms of origin, type, and severity, rather than random output error alone.

The taxonomy supports two primary uses. First, it enables stage localization by identifying which component introduced a fault into the pipeline. Second, it supports targeted mitigation by distinguishing faults that can be repaired through rule adjustment from faults that require more advanced capabilities, such as explicit hierarchy-aware header representation or support for multiple regions and region selection.

TABLE II. Failure Taxonomy

Code	Stage	Failure Type	Symptom
FA2	Ingestion	Merge mis-resolution	Merged spans expanded incorrectly, and hierarchy spans were lost
FB1	Region detection	Under-selection	Table region clipped and misses rows or columns
FB2	Region detection	Over-selection	Notes or adjacent blocks included as part of the table
FC1	Structural interpretation	Header depth error	Wrong number of header rows or columns inferred
FC2	Structural interpretation	Boundary ambiguity	Header-body boundary misidentified due to mixed types
FC3	Structural interpretation	Multi-region confusion	The wrong table was selected when multiple tables exist
FD1	Semantic labeling	Role mislabeling	Headers treated as data or notes treated as fields
FD2	Semantic labeling	Hierarchy collapse	Parent-child header relations lost during flattening
FE1	Canonicalization	Wrong granularity	Subtotals or group headers are emitted as observations
FE2	Canonicalization	Key construction failure	Composite keys inconsistent or missing
FF1	Validation	Silent semantic error	Output passes schema checks but is semantically wrong

VI. Failure Matrix and Analysis

The pipeline (R1-R6) was developed in Java using Apache POI for XLSX file handling, including access to merged-cell metadata. Each controlled case (S1-S6) was processed using the same rule-execution settings. Therefore, differences in behaviour are attributable to layout variation rather than rule configuration or implementation changes. During rule execution, the pipeline logs stage-visible failure events. Each event is assigned a taxonomy code (FA2-FF1) and a severity score from 0 to 2, as described in Section IV-D. Table III is constructed by collecting logs from all controlled cases and reporting the maximum severity for each pair of controlled case and taxonomy code. The raw logs were exported in CSV format to support reproducibility and future extension of the rule-based execution process [7], [5].

TABLE III. Failure Matrix

Failure Code	S1	S2	S3	S4	S5	S6
FA2	0	2	0	2	0	2
FB1	0	0	0	0	0	0
FB2	0	0	0	0	1	1
FC1	1	0	1	0	1	1
FC2	0	0	2	0	0	2
FC3	0	0	2	0	2	2
FD1	0	0	2	0	2	0
FD2	1	1	1	1	0	1
FE1	0	0	0	0	2	2
FE2	0	0	0	0	0	0
FF1	0	0	2	0	2	2

For consistent reporting, we generated a matrix that maps taxonomy codes (Table II) to the specific controlled cases (S1-S6). Table III reports measured severity levels ranging from 0 to 2, where a level of 0 represents a successfully covered case. Severity scores were derived by executing the baseline rule pipeline (R1-R6) against each of the test cases, logging each triggered failure according to the procedure discussed in Section IV-D. For every combination of (case, taxonomy-code) that triggered a failure, the matrix stores the highest severity observed among the triggering checks.

Observed patterns. The matrix highlights three predominant failure clusters:

1) Merged header and hierarchical issues (FA2, FD2). A set of sheets exhibits hierarchical or merged header formatting (S2, S4, and S6) that results in consistent FA2 (severity 2) and FD2 (severity 1). Expanding a merged header and subsequent flattening of the header hierarchy preserves the ability to parse structure, but results in incorrect parent-child header relations.

2) Multi-block and boundary problems (FC2, FC3). The collection of mixed-structure worksheets (S3, S5, and S6) consistently yields FC3 (severity 2), representing ambiguity when several blocks, borders, or sections coexist on a single sheet. For S3 and S6, this also comes with FC2 (severity 2) because determining header, body separation breaks down when layout hints, and data types are interwoven (including section title entries, repeated headers, and values appearing at or near the header rows).

3) Semantic corruption and blind errors (FD1, FE1, FF1). When mixed-structure patterns are encountered, errors can propagate down to later analysis stages. The cases S3 and S5 incur FD1 (severity 2), failing to correctly attribute the roles of headers that appear in duplicate sections and data rows that have been identified as part of the header. Cases S5 and S6 receive FE1 (severity 2), showing incorrect cardinality resulting from rows formatted as subtotals and various summarized values or artifacts being included as raw observations. These failures frequently result in FF1 (severity 2) for S3, S5, and S6; although a portion of the table might appear correctly formatted, the extracted semantic structure is flawed.

Cases with less impact. The only case that shows mild consequences is S1 (which has a single tidy table arrangement), generating a minor header depth problem (FC1=1) and an indicator of incomplete hierarchical parsing (FD2=1), which is indicative that such simple contiguous tables are correctly handled by the parser's baseline mode. For cases S2 and S4, the main failures are found in the header structure hierarchy handling (FA2/FD2) and not in downstream semantic extraction.

Code occurrences. There were zero occurrences of FB1 and FE2 codes in the experiments under the baseline configuration of the table extractor. FB1 is rare because the extractor's default bounding behaviour would ensure all relevant non-empty cells are within the defined scope, FE2 only triggers in the context of user-defined, multipart keys being constructed, and stronger tests for key consistency than are required to extract Table III in this study.

VII. Implications and Recommendations

The analysis in Table III indicates that more than half of the observed failures stem from errors that commonly arise in patterns of spreadsheets rather than from a deficiency in a rule covering a rare instance. This situation indicates the necessity to substantially extend the capabilities of the transformation pipeline to handle widespread errors more effectively. The common failure categories were used in our analysis to recommend actions for each stage of the transformation pipeline (stages A through F) according to our taxonomical categories (FA2 through FF1).

A. Explicit Representation of Header Hierarchy (FA2, FD2)

The cases that utilize merged or hierarchical headers (S2, S4, S6) often correlate to misinterpretation of merge resolution (FA2) and collapse of hierarchy (FD2). The method by which we treat merged cells cannot be limited to mere preprocessing. A hierarchical tree or graph that captures the header information should store parent-child relationships between header cells and provide the appropriate content when flattening occurs. The flattening logic should produce consistent, determined outcomes, but this process requires knowledge of the header hierarchy. Merging cells should occur only after establishing a structural hierarchy. For example, "2023" (Level 2) should precede "Revenue" (Level 1), yielding "2023 - Revenue" rather than "Revenue - 2023", thus removing ambiguity and duplicated "Revenue" headers.

B. Multi-Block Reasoning Before Header Inference (FC2, FC3)

Misinterpreting regions (FC3) typically arises in our mixed layout datasets (S3, S5, S6) and is associated with ambiguous region boundaries (FC2). Header and table selection should not be restricted based on only a single bounding box. Often, worksheets contain irrelevant notes or extra separators or consist of multiple independent blocks of information. We add a distinct segmentation phase that identifies dense regions visually by looking for many empty rows, columns, or explicit section separators. For each such region, we assess how likely it is to contain a table before we run depth-based header heuristics. Section labels like "Class A" and "Class B" should not be treated as headers; rather, they should be treated as meta information to the table situated underneath them, to ensure that when headers are flattened, a string such as "Class A" in a data row does not accidentally become a column header.

C. Enforcement of Role Labeling and Granularity Control (FD1, FE1)

Mislabelling the roles of cells (FD1) and issues in granularity (FE1) tend to happen with recurring headers and subtotal/total-like rows in S3, S5, and S6. If the parser treats these repeating structures as a data row, it mislabels roles and proceeds to incur granularity errors. Two principal controls are proposed for addressing such problems: first, identify repeated headers by detecting matches between the inferred content of a header row and some data within a table region and second, introduce a more stringent granularity filter capable of

distinguishing true data observations from summary data points, such as total, subtotal, average, etc., based on keyword identification as well as structure, formatting and the density of numbers.

D. Semantic Invariant Checks for Silent Failure Detection (FF1)

Silent semantic errors (FF1), where the output is still tabular but meaningless in its content (S3, S5, S6), imply that we must supplement our Stage F validation logic with semantic invariant checks in addition to validating the presence of a schema and data types. Such invariants can be used to: detect incomplete keys (by checking whether key-like columns possess relatively few missing values), define the range of the number of records within each section (e.g., within a 10-20% variation based on prior inspection), verify the match between the inferred section labels and the data within those sections and find any header labels that happen to appear inside the data portion of a table. All these checks can be implemented easily without any ML or deep learning model, and can go a long way toward avoiding incorrect outputs.

E. Summary: Local Rule Tuning Versus Capability Enhancements

Based on the structure of the failure matrix, we have found that a large portion of the errors fall into two distinct types: failures resolvable by fine-tuning existing rules (e.g., expanding keyword sets, modifying numerical thresholds) and failures requiring significant improvements in the pipeline capability (e.g., header representation that handles hierarchy, block-level segmentation). Successfully increasing the robustness of spreadsheet parsers in production systems requires a combined approach of: representation of explicit structure (header spans, block segmentation), better controls for cell role and granularity, and semantic invariant checks to prevent silent misrepresentation of data.

VIII. Threats to Validity

The findings are tied to a specific deterministic baseline rule configuration, R1-R6. The specific events that trigger failures may shift under a different sequence of pipeline stages or under different rule parameters. For example, changing a header-depth heuristic or keyword list could cause a failure to be reported at one stage rather than another. Therefore, to keep results comparable, the rules set must remain consistent and deterministic across all inputs. The matrix in Table III reports the maximum severity score for each case-code pair, ensuring that a triggered failure is not hidden if it appears earlier in the pipeline.

A. Construct Validity

The taxonomy codes (FA2–FF1) and severity scores (0–2) defined in this study provide an operational definition of what is treated as a table-structure error. However, spreadsheet cells may be interpreted differently by different users because their meaning is often shaped by layout, domain context, and the intended use of the worksheet. In the controlled cases, each cell was treated according to an explicit row-and-column interpretation so that the expected output could be evaluated consistently. Nevertheless, some spreadsheet entries may have ambiguous semantic roles. For example, a date-like entry may be interpreted either as a section note or as a data value depending on the user's expectation and domain context. One user may regard such a row as a contextual notice that should be ignored, whereas another user may regard it as part of the data that should be parsed. Therefore, the failures reported in this study are limited to the explicitly checked conditions defined in the evaluation procedure. The severity scores should also be interpreted according to the adopted rubric: a score of 1 indicates a failure that can be repaired through localized adjustment of existing rule parameters, such as threshold tuning or keyword refinement, whereas a score of 2 indicates a missing pipeline capability that requires additional functionality, such as explicit hierarchy representation or multi-block reasoning across related table regions.

B. External Validity

This study considers six common forms of data organization: simple tidy tables, tables with multi-row headers, sectioned blocks, a mixture of structures, report style layout, and a header and then region-level nested lists. However, we do not know the real distribution of types of tables, and this study is only concerned with tables that adhere to these patterns, for instance, more formal types like PivotTable structures, which include formulas that span arrays, multi-sheet workbooks whose information might be pulled from external locations, or other spreadsheets whose formatting goes beyond mere style, are unexamined. Therefore, the pattern of observed failures in Table III should not be inferred as representative statistics of spreadsheet errors in all circumstances; rather, it should be seen as an illustration of fragile components in spreadsheet structures. This future study could further investigate and expand our work through more diverse test data: more domains and actual spreadsheets that include varied structures.

C. Reliability and Reproducibility

The correctness and ability to repeat this process may be affected by nuances of the implementation details, including variations between libraries, different ways of handling, and unified formats of merged cells (which can be consistent throughout various spreadsheet editors but still represent variability between the underlying editors), and different interpretations of spreadsheet data when normalized. As such, the evaluation process is structured to be deterministic with a controlled, baseline system that details and outputs errors that Table III summarizes under an unambiguous reporting of a failure type (maximum per cell code per case). With other rule-based or data parsers, different findings might arise because other techniques or assumptions were applied, especially concerning heuristics used in region segmentation or identification of cell headers of certain types, in addition to possible variation in algorithm implementations of parsing merged cells. Regardless, these findings provide a systematic and reproducible way to categorize, evaluate, and situate failure modes for conversion processes of spreadsheets to tabular form.

IX. Conclusion and Future Work

This study introduces a failure analysis framework for rule-based spreadsheet-to-relational transformations. The framework uses a stage-based taxonomy spanning ingestion,

region detection, structural interpretation, semantic labeling, canonicalization, and validation. The taxonomy enables failures to be attributed to specific causes and processing stages. By controlling six spreadsheet cases (S1-S6) and using a fixed baseline implementation, the study constructs a failure matrix that records the maximum severity for each case-code pair. The results show that failures are concentrated in three groups: 1) handling merged cells and hierarchical headers (FA2, FD2), 2) handling separate table blocks and ambiguous boundaries (FC2, FC3), and 3) downstream semantic errors and wrong granularities that can lead to silent data corruption (FD1, FE1, FF1). These findings indicate that many errors arise not from trivial rule mistakes but from the inability of extraction systems to represent and enforce explicit structural models.

Future work will extend the controlled spreadsheet cases in two directions: by including more diverse real-world spreadsheets from different domains and by evaluating alternative baseline implementations to determine whether the observed issues can be mitigated through rule adjustment or whether they stem from deeper system constraints. This broader evaluation is consistent with recommendations in [5], which emphasize spreadsheet extraction pipelines that integrate recognition, interpretation, data cleaning, and evaluation across different table formats. Subsequent mitigation work will focus on three areas motivated by the most prevalent failure types in Table III: designing header models that preserve merged-cell and parent-child relationships without reducing them to single-level headers, performing table-block segmentation and ranking before header analysis, and introducing semantic invariant checks such as key completeness, section consistency, and value reconciliation to prevent silent errors. Finally, the taxonomy and logging tools will be released as a benchmark kit so that other rule-based systems can report failures uniformly, compare mitigation strategies, and demonstrate which improvements address which failure codes. In the longer term, the work will move beyond failure analysis toward the semantics of canonical tables, such as linking extracted values to external knowledge resources after the underlying structural and semantic failures have been substantially reduced [9].

REFERENCES

[1] Tomas Petricek, Gerrit J. J. van den Burg, Alfredo Nazabal, Taha Ceritli, Ernesto Jimenez Ruiz, and Christopher K. I. Williams, “AI Assistants A Framework for Semi-automated Data Wrangling,” IEEE Transactions on Knowledge and Data Engineering, Vol. 35, No. 9, 2023, pp. 9295–9306.

[2] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer, “Wrangler Interactive Visual Specification of Data Transformation Scripts,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 3363–3372.

[3] Sumit Gulwani, William R. Harris, and Rishabh Singh, “Spreadsheet Data Manipulation Using Examples,” Communications of the ACM, Vol. 55, No. 8, August 2012, pp. 97–105.

[4] Alexey O. Shigarov and Andrey A. Mikhailov, “Rule-Based Spreadsheet Data Transformation from Arbitrary–Relational Tables,” Information Systems, Vol. 71, 2017, pp. 123–136.

[5] Alexey O. Shigarov, Vasiliy V. Khristyuk, Viacheslav V. Paramonov, Alexandr Yu. Yurin, and Nikita O. Dorodnykh, “Toward Framework for Development of Spreadsheet Data Extraction Systems,” CEUR Workshop Proceedings, Vol. 2221, 2018, pp. 1–7.

[6] Alexey O. Shigarov, “Table Understanding Using a Rule Engine,” Expert Systems with Applications, Vol. 42, No. 2, 2015, pp. 929–937.

[7] A. Shigarov, V. Khristyuk, and A. Mikhailov, “TabbyXL Software Platform for Rule-Based Spreadsheet Data Extraction and Transformation,” SoftwareX, Vol. 10, 2019, Article 100270.

[8] A. Shigarov, “Rule-Based Table Analysis and Interpretation,” in Information and Software Technologies, 21–22 October 2015, Kaunas, Lithuania, Communications in Computer and Information Science, Vol. 538, Springer, 2015, pp. 175–186.

[9] Nikita O. Dorodnykh and Aleksandr Yu. Yurin, “Towards a universal Approach for Semantic Interpretation of Spreadsheets Data,” Proceedings of the 24th International Database Engineering and Applications Symposium, 12–14 August 2020, Seoul, Republic of Korea, 2020, 9 pages.

[10] Sara Bonfitto, Elena Casiraghi, and Marco Mesiti, “Table Understanding Approaches for Extracting Knowledge from Heterogeneous Tables,” WIREs Data Mining and Knowledge Discovery, 2021.

[11] Majid Ghasemi Gol, Jay Pujara, and Pedro A. Szekely, “Tabular Cell Classification Using Pre-trained Cell Embeddings,” Proceedings of the 2019 IEEE International Conference on Data Mining, 2019.

[12] P. Yin, G. Neubig, W. Yih, and S. Riedel, “TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8413–8426.

[13] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: Table Understanding through Representation Learning,” Proceedings of the VLDB Endowment, Vol. 14, No. 3, 2020, pp. 307–319.

[14] Z. Wang, H. Dong, R. Jia, J. Li, Z. Fu, S. Han, and D. Zhang, “TUTA: Tree-Based Transformers for Generally Structured Table Pre-training,” Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1780–1790.

[15] H. Iida, D. Thai, V. Manjunatha, and M. Iyyer, “TABBIE: Pretrained Representations of Tabular Data,” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 3446–3456.

[16] G. Badaro, M. Saeed, and P. Papotti, “Transformers for Tabular Data Representation: A Survey of Models and Applications,” Transactions of the Association for Computational Linguistics, Vol. 11, 2023, pp. 227–249.

Arabic Sign Language Alphabet Recognition Using Transfer Learning: Evaluation, Ablation, and Deployment

Abdelfatah Maarouf[1], Otman Maarouf[2], Abdelaali Benaiss[3], Rachid El Ayachi[4], Mohamed Biniz[5]
Department of Computer Science-Faculty of Science and Technology,
Laboratory TIAD, Sultan Moulay Slimane University, BP 523, Beni Mellal, 23000, Morocco[1, 2, 3, 4]
Centre of Excellence-Faculty of Sciences-Laboratory LabSIV, Ibnou Zohr University, Agadir, 80000, Morocco[2]
Department of Computer Science-Polydisciplinary Faculty Laboratory LIMATI,
Sultan Moulay Slimane University, BP 592, Beni Mellal, 23000, Morocco[5]

***Abstract*—Arabic Sign Language (ArSL) is one of the most widely used sign languages among the hearing-impaired community in Arabic-speaking regions. Yet, the automated recognition of its alphabet remains a critical challenge for assistive technology development. This study presents a transfer learning classification model for Arabic Sign Language Alphabets (ArSLA) based on the InceptionV3 convolutional neural network architecture. The research contribution is a rigorously evaluated, reproducible recognition pipeline that advances diagnostic depth beyond prior studies through ablation testing, per-class F1-score analysis, and confusion pattern interpretation. The ArSL2018 dataset, comprising 54,049 images distributed across 32 Arabic alphabet classes, was used for training and evaluation. The model integrates transfer learning from ImageNet-pretrained weights, Global Average Pooling, a 256-unit dense layer with ReLU activation, dropout regularization (rate: 0.4), and a 32-class softmax output layer. Training employed the Adam optimizer with adaptive learning rate scheduling and early stopping callbacks. Model evaluation was conducted using a stratified 80:20 train-test split replicated across five independent runs with different random seeds, yielding a mean test accuracy of 97.41% ± 0.31% and a best single-run test accuracy of 97.68%, outperforming all previously reported models on the same benchmark dataset. An ablation study confirmed the independent contributions of transfer learning, data augmentation, and dropout regularization. A real-time prototype was implemented using OpenCV at 3.06 FPS on CPU hardware. These findings establish InceptionV3-based transfer learning as a strong and reproducible baseline for Arabic sign language assistive technology.**

Keywords*—Arabic Sign language alphabets; deep learning; transfer learning; inceptionv3; ArSL2018; gesture recognition; hearing impairment

I. Introduction

Sign languages are the primary medium of communication for Deaf communities worldwide, yet automated tools capable of interpreting them accurately remain far from ubiquitous. For the roughly 70 million people globally who are Deaf or hard of hearing, the absence of reliable machine translation [1] of signed messages constitute a genuine barrier to participation in everyday life. Arabic Sign Language (ArSL) is the dominant signed modality across Arab-majority countries [2], and within it, the alphabet-based subset Arabic Sign Language Alphabets (ArSLA) offers a tractable entry point for machine recognition [3], [4], [5] because each of the 32 classes corresponds to a single static hand configuration. Building a robust classifier for these gestures is, therefore, a concrete and socially meaningful objective [6].

What makes ArSLA harder to recognise than it might initially appear is the fine-grained nature of its visual distinctions. Unlike the broad, whole-word signs of, say, American or British Sign Language, Arabic alphabetic gestures rely on subtle cues: the precise angle of a curled finger, the placement of the thumb relative to the palm, and the spacing between digits that in the Arabic script would correspond to diacritical marks. Generic deep learning architectures encounter characteristic difficulties here. Standard ResNet models [7], designed for coarse object-level discrimination, tend to blur over these fine inter-class differences unless carefully fine-tuned for the task. Skeleton-based MediaPipe pipelines discard the very texture and shape detail that distinguishes similar-looking letters [8]. Vision Transformers [9], powerful as they are in large-data regimes, require considerably more training data than ArSL2018 provides before they converge reliably. InceptionV3 [10] sidesteps several of these limitations: its parallel convolutional branches capture features at multiple spatial scales simultaneously, which is well-suited to the dual need for global hand-shape context and localised finger-joint detail.

Prior work on ArSLA has grown steadily over the past decade. Alamri and Lajmi [11] built a real-time translation system around a fine-tuned SSD-ResNet50 V1 FPN pipeline, achieving 94% accuracy. Ismail et al. applied both fused 2D/3D CNN-recurrent models for dynamic ArSL [12] and DenseNet121 together with VGG16 for static recognition on a 220,000-image corpus [13]. Alani and Cosma [14] trained their custom ArSL-CNN directly on the ArSL2018 benchmark after addressing class imbalance with SMOTE, reaching 97.29%. More recently, Zakariah et al. [15] applied EfficientNet-based transfer learning to the same dataset and reported 95.0%, while Alsaadi et al. [16] obtained 94.81% with AlexNet. Each of these contributions advances the field, yet several important gaps persist.

None of the studies above reports results across multiple independent experimental runs, leaving open the question of result stability. Ablation evidence isolating which components of a recognition pipeline actually matter is similarly absent from the ArSL2018 literature. Per-class diagnostic metrics and confusion pattern analysis, essential for understanding where a model fails and why, have received only superficial treatment. The present work addresses all three of these gaps directly.

The research contribution of this work is as follows: 1) A transfer learning pipeline based on InceptionV3, evaluated on the ArSL2018 benchmark, and achieving a mean test accuracy of 97.41% ± 0.31% across five independent runs, the highest reported figure for this dataset, with statistical significance confirmed against the nearest competitor. 2) A reproducible five-run evaluation protocol that replaces single-split reporting with mean and standard deviation across varied random seeds. 3) An ablation study that quantifies the separate contributions of transfer learning initialisation, data augmentation, and dropout regularisation to final test accuracy. 4) A per-class F1-score analysis paired with a systematic interpretation of the confusion matrix, identifying the specific sign pairs responsible for most misclassifications and explaining the underlying geometric reasons. 5) A real-time OpenCV prototype that translates the trained model into a functional assistive tool, providing an honest deployment baseline including inference speed measured on commodity hardware.

The remainder of this study is structured as follows. Section II reviews the relevant literature on sign language recognition, deep learning for gesture classification, and Arabic-specific studies. Section III describes the proposed methodology in full. Section IV presents the experimental results and discusses their implications. Section V concludes the study with a frank assessment of limitations and a set of concrete directions for future research.

II. Related Work

The literature on automated sign language recognition [17], [18] spans sensor-based, skeleton-based, and vision-based approaches, with deep learning having become the dominant paradigm over the past five years. The following subsections survey key contributions across three thematic areas most relevant to the present work.

A. Sign Language Recognition Models

Sensor-based methods were among the earliest to achieve competitive accuracy. Amin et al. [19] fitted hearing-impaired volunteers with an E-Voice smart glove instrumented with flex and inertial sensors, obtaining reliable recognition of American Sign Language alphabets in real time, though the approach requires users to wear specialised hardware. Vision-based methods quickly demonstrated that cameras alone can be sufficient. Shah et al. [20] tackled Pakistan Sign Language with a multiple kernel learning framework that combined hand-crafted shape descriptors with learned features, showing that language-specific feature engineering yields marked improvements over generic classifiers applied off the shelf. Attention mechanisms opened another productive avenue: Pan et al. [21] incorporated keyframe sampling and skeletal attention into a recognition network, demonstrating that selectively weighting the most informative frames substantially reduces confusion among visually similar signs. Alongside these static-image and video approaches, Zhao et al. [22] showed that low-cost wearable devices can serve as credible alternatives to optical cameras for gesture recognition when paired with appropriately designed models. For continuous, sentence-level translation, Papastratis et al. [23] proposed cross-modal alignment of video and text embeddings in a shared latent space, pointing toward the more ambitious goal of full sign-language sentence understanding rather than isolated letter classification.

B. Deep Learning in Sign Language Classification

The adoption of deep convolutional neural networks [24], [25], [26], [27] transformed gesture recognition much as it transformed general computer vision. Alasmari and Asiri [28] applied ResNet [6] to sign language classification and found that the skip connections, which make ResNet so effective for object recognition, transfer reasonably well to hand gesture discrimination, provided the base model is fine-tuned on gesture-specific data rather than used purely as a frozen feature extractor. Buttar et al. [29] pushed this further by proposing a hybrid deep learning architecture that handles both static and dynamic signs within a single model, arguing that the two sub-problems are better addressed jointly than in isolation. Al-Qurishi et al. [30] contributed an extensive survey of the field, cataloguing benchmark datasets, evaluation protocols, and open challenges; their analysis is particularly relevant here because it highlights cross-dataset generalisation and real-time deployment as the two most pressing unsolved problems. On the deployment side, Breland et al. [31] demonstrated that thermal-image-based sign digit recognition can run on edge computing hardware with acceptable latency, an important existence proof for resource-constrained scenarios. Al-Hammadi et al. [32] showed that the choice of hand gesture representation feeding a deep network matters significantly: carefully engineered spatial descriptors consistently outperform naive pixel-level inputs, even when the downstream classifier is identical.

C. Arabic Sign Language Studies

Research focused specifically on Arabic Sign Language has grown in volume and sophistication. Ibrahim et al. [33] developed one of the earliest automatic ArSL recognition systems, relying on hand-crafted geometric features and traditional classifiers; while the accuracy now appears modest by current standards, their work established the experimental conventions, dataset structure, class definitions, and evaluation metrics that later benchmark efforts built upon. Deriche et al. [34] took a different path, using a pair of Leap Motion Controllers in conjunction with Gaussian Mixture Model classification to capture three-dimensional hand pose data for 32 Arabic letter classes, reporting that dual-sensor input substantially reduced the ambiguity that plagues single-camera systems for letters with near-identical 2D projections. Tharwat et al. [35] provided a systematic comparison of several classical machine learning classifiers, support vector machines, k-nearest neighbours, and random forests on the same alphabet recognition task, offering a useful pre-deep-learning baseline against which subsequent CNN-based [36] results can be contextualised. The transfer learning surveys of Zhuang et al. [37] and Pan and Yang [38] contextualise why ImageNet pretraining transfers well to domain-specific recognition tasks:

low-level edge and texture detectors learned on millions of natural images remain broadly useful even when the target domain differs substantially from the source. Alsaadi et al. [16] and Zakariah et al. [15] both applied transfer learning to the ArSL2018 benchmark, with AlexNet and EfficientNet, respectively, establishing the quantitative baselines against which the present study is compared. The closest prior result is that of Alani and Cosma [14], whose ArSL-CNN achieved 97.29% on the same dataset.

Taken together, this body of work leaves three notable gaps. No study on ArSL2018 has reported results over multiple independent experimental runs, assessed the contribution of individual pipeline components through ablation, or provided a systematic per-class error analysis explaining the geometric reasons behind misclassifications. Each of these gaps is addressed in what follows.

III. METHODS

The experimental pipeline is summarised in Fig. 1 and proceeds through five main stages: dataset preparation, preprocessing and augmentation, model construction, training, and evaluation. Each stage is described in detail below. Throughout, the emphasis is on transparency and reproducibility; every design choice is motivated, every parameter is reported, and the evaluation protocol is explicitly designed to produce statistically reliable estimates rather than single-run figures.

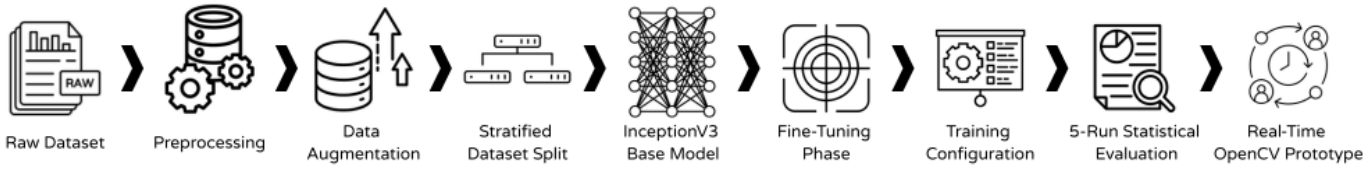


Fig. 1. Overview of the proposed research pipeline.

A. Dataset

All experiments use ArSL2018 [39], a publicly available benchmark designed specifically for Arabic Sign Language alphabet recognition. The dataset contains 54,049 images distributed across 32 classes, one for each letter of the Arabic alphabet, with an average of approximately 1,689 images per class (standard deviation: ±142 images). This near-uniform class distribution removes class imbalance as a confounding factor, so all reported metrics reflect genuine per-class recognition difficulty rather than artefacts of skewed training data.

Images were collected at Prince Mohammad Bin Fahd University in Saudi Arabia, with 40 volunteers of varying ages signing against diverse backgrounds at a fixed camera distance of approximately one metre. The resulting variation in skin tone, lighting, and background texture gives the dataset a degree of natural diversity that purely controlled studio datasets lack, though it remains a controlled environment rather than a fully unconstrained real-world collection.

Table I presents the dataset classes along with their corresponding indices and labels, with all 32 classes well-represented. Representative samples from the dataset are shown in Fig. 2.

TABLE I. CLASS INDEX AND LABEL

Index	Class Label	Index	Class Label
0	Ain	16	Laam
1	Al	17	Meem
2	Aleff	18	Nun
3	Bb	19	Ra
4	Dal	20	Saad
5	Dha	21	Seen
6	Dhad	22	Sheen
7	Fa	23	Ta
8	Gaaf	24	Taa
9	Ghain	25	Thaa
10	Ha	26	Thal
11	Haa	27	Toot
12	Jeem	28	Waw
13	Kaaf	29	Ya
14	Khaa	30	Yaa
15	La	31	Zay

Fig. 2. Representative samples from the ArSL2018 dataset.

B. Preprocessing and Augmentation

All images were resized to 150×150 pixels and converted to the [0, 1] floating-point range via division by 255. InceptionV3

was originally designed to accept 299×299 inputs, but preliminary experiments on a held-out validation subset showed that 150×150 images yield comparable classification accuracy while reducing training time by roughly 40% on the available CPU hardware, a worthwhile trade-off given the deployment context where inference speed is a practical constraint.

The dataset was partitioned into training (80%) and test (20%) subsets using stratified sampling, so each class is represented in both splits in proportion to its overall frequency. This split was applied consistently across all five experimental runs by setting the random seed before sampling; the seed values were 42, 7, 21, 99, and 2024, respectively. Augmentation was applied exclusively to the training set using TensorFlow's ImageDataGenerator with two transformations: random zoom up to 20% and horizontal flipping. These choices were deliberate. Zoom variation simulates signers at different distances from the camera, a realistic source of intra-class variation in deployed systems. Rotation augmentation, by contrast, was intentionally excluded: because angular hand orientation is semantically meaningful in ArSLA (rotating a sign can change which letter it represents), applying arbitrary rotations to training images would introduce linguistically invalid samples and likely harm rather than help generalisation. Augmented training samples are illustrated in Fig. 3.

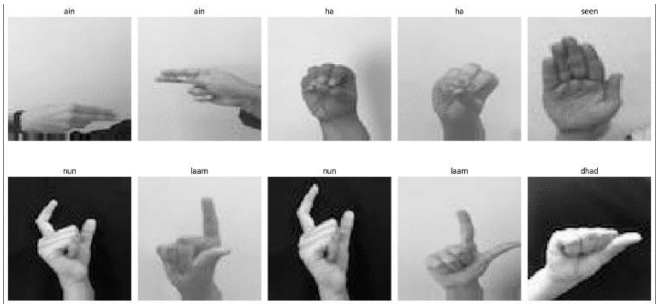


Fig. 3. Examples of augmented training images.

C. Model Architecture

The backbone of the proposed model is InceptionV3 [10], loaded with weights pre-trained on ImageNet [40] and with the original top classification layer removed. The choice of InceptionV3 over other commonly used architectures was guided by three considerations specific to this task. First, the Inception module's parallel convolutional branches operating at 1×1, 3×3, and 5×5 kernel sizes simultaneously extract features at multiple spatial scales in a single pass, which suits the ArSLA recognition problem well: discriminating between letter classes requires both global awareness of overall hand shape and fine-grained sensitivity to individual finger positions. Second, InceptionV3's factorised convolutions keep parameter count lower than comparably deep ResNet variants [7], reducing the risk of overfitting when fine-tuning on a domain-specific dataset of moderate size. Third, unlike EfficientNet's compound scaling strategy [41], InceptionV3 tends to exhibit stable gradient flow during fine-tuning, even at modest batch sizes, a practical advantage when GPU acceleration is unavailable. The overall architecture is depicted in Fig. 4.

Four layers were stacked on top of the base model to form the classification head: a Global Average Pooling layer, which collapses the spatial dimensions of the final feature map and reduces the risk of spatial overfitting; a fully connected layer of 256 units with ReLU activation; a Dropout layer with a rate of 0.4 [42], applied during training to prevent co-adaptation of feature detectors; and a softmax output layer with 32 units corresponding to the 32 ArSLA classes. The complete model architecture is depicted in Fig. 5.

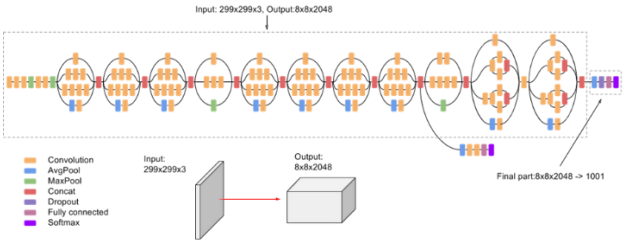


Fig. 4. The inceptionV3 architecture.

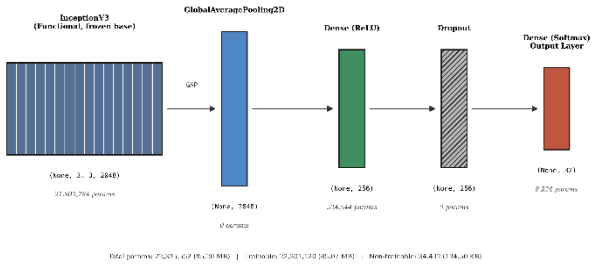


Fig. 5. Complete architecture of the proposed model.

Because the model uses a softmax output layer, it assigns a probability to every class regardless of the input. In practice, this means that a hand pose not belonging to any of the 32 ArSLA classes, an out-of-distribution (OOD) gesture, will still receive a confident-looking prediction. To mitigate this in the real-time prototype, a confidence threshold of 0.85 was applied: any frame for which no class achieves a softmax score above this value is labelled "Unknown Gesture" rather than assigned the top-scoring class. This is not a rigorous OOD detection solution, energy-based scoring [43] or temperature scaling would offer stronger statistical guarantees, but it provides a practical safeguard against obviously spurious predictions in the prototype context.

D. Training Protocol

Training proceeded in two phases to make use of the pre-trained weights while allowing meaningful task-specific adaptation. In Phase 1, all 311 layers of the InceptionV3 base were frozen; only the custom head was trained, allowing the classification layers to find sensible initial weights without corrupting the general-purpose representations learned from ImageNet. From epoch 10 onward, Phase 2 began: the last 50 layers of the base model spanning the mixed9 inception block were unfrozen and trained jointly with the head at the reduced learning rate supplied by the scheduler. This two-phase approach is a widely used fine-tuning strategy that balances retention of transferable low-level features with adaptation of higher-level representations to the target domain.

The full set of hyperparameters is recorded in Table II. The Adam optimizer [44] was selected for its adaptive per-parameter

learning rates, which generally accelerate convergence relative to standard stochastic gradient descent. Three Keras callbacks governed training: ModelCheckpoint saved the weights that achieved the best validation accuracy; EarlyStopping halted training if validation loss showed no improvement for five consecutive epochs; and ReduceLROnPlateau reduced the learning rate by a factor of 0.1 whenever validation loss plateaued for three consecutive epochs, enabling finer parameter updates as training progressed.

TABLE II. SUMMARY OF HYPERPARAMETERS AND TRAINING CONFIGURATION

Parameter	Value
Base model	InceptionV3 (ImageNet weights)
Frozen layers: Phase 1	All 311 InceptionV3 layers
Fine-tuned layers: Phase 2	Last 50 layers (mixed9 block onward, from epoch 10)
Classification head	GAP → Dense(256, ReLU) → Dropout(0.4) → Dense(32, Softmax)
Optimiser	Adam [44]
Initial learning rate	0.001 (minimum: 1×10^{-6})
Loss function	Categorical cross-entropy
Batch size	32
Maximum epochs	40
Input image size	150×150 pixels, RGB
Random seeds (five runs)	42, 7, 21, 99, 2024
Hardware	Intel Core i5-3427U @ 1.80 GHz, 8 GB RAM, CPU only
Software	Python 3.9, TensorFlow 2.12, Keras, OpenCV 4.7, scikit-learn 1.2

To generate statistically meaningful performance estimates, the entire training and evaluation procedure was repeated five times under different random seeds. Each seed controls weight initialisation, batch shuffling, and the stratified split, so the five runs are genuinely independent realisations of the same experiment rather than minor perturbations of a single run. Reporting mean and standard deviation across these runs is a conservative safeguard against overstating the reliability of results, a particularly important consideration given that single-split evaluations of this kind are known to exhibit non-negligible variance even on well-balanced datasets [45].

IV. RESULTS AND DISCUSSION

The following subsections present the experimental results in a deliberate order: training dynamics first to characterise the learning process, then aggregate quantitative metrics, then the ablation study, then per-class and confusion analysis, then comparison with published baselines, and finally a discussion of inference speed, practical implications, and the honest limitations of the work.

A. Training Dynamics

Fig. 6 shows the accuracy curves for a representative run (seed 42). Training accuracy climbed steadily to 99.5%, while validation accuracy tracked closely behind, settling at 97.68%. The narrow gap between the two throughout training is reassuring: it suggests that the combination of dropout regularisation, moderate augmentation, and the two-phase fine-tuning schedule successfully prevented the network from memorising the training set at the expense of generalisation. No abrupt divergence between the curves, the hallmark of overfitting, appeared in any of the five runs.

The corresponding loss curves, shown in Fig. 7, tell a consistent story. Both training and validation loss declined smoothly and converged without the spiking or plateauing patterns that indicate learning rate misspecification. The ReduceLROnPlateau callback triggered on six occasions, each time nudging the optimiser toward finer updates; these events are visible as small inflections in the validation loss curve.

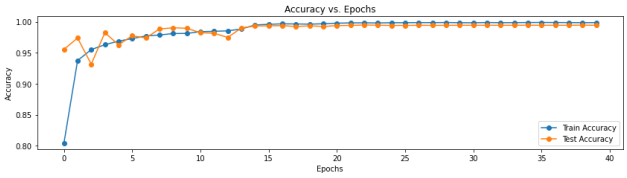


Fig. 6. Training and validation accuracy across epochs.

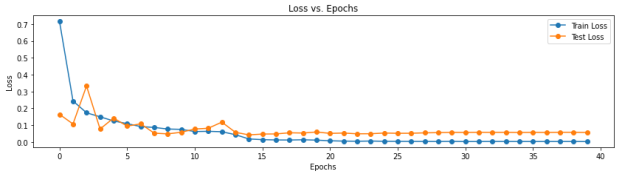


Fig. 7. Training and validation loss across epochs.

Fig. 8 explicitly plots the learning rate schedule. Starting at 1×10^{-3}, the rate was reduced six times by a factor of 0.1, reaching 3.16×10^{-6} by the final epochs. This graduated decay allowed the model to take large steps early in training when the parameters were still far from a good solution, and progressively smaller steps later when the task was to fine-tune within a narrow region of the loss landscape rather than explore broadly.

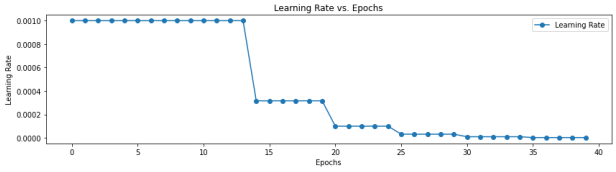


Fig. 8. Learning rate schedule applied during training.

B. Quantitative Performance

Table III reports precision, recall, F1-score, and accuracy for each of the five runs, together with the mean and standard deviation across runs. The standard deviations are consistently below 0.4 percentage points, confirming that the reported figures are not sensitive to the particular random seed used and can be trusted as representative estimates of the model's true performance on the ArSL2018 distribution. The best single run (seed 42) achieved 97.68% test accuracy; the mean across all five runs was 97.41% ± 0.31%.

TABLE III. PERFORMANCE METRICS ACROSS FIVE INDEPENDENT EXPERIMENTAL RUNS

Metric	Training Set	Best Run (Seed 42)	Mean ± Std (5 Runs)
Accuracy	99.5%	97.68%	97.41% ± 0.31%
Precision	99.61%	98.02%	97.78% ± 0.28%
Recall	99.54%	97.81%	97.58% ± 0.33%
F1-score	99.57%	97.91%	97.68% ± 0.30%

C. Ablation Study

To understand how much each component of the pipeline contributes to the final result, four configurations were trained and evaluated under the same conditions (seed 42, 80/20 split). The results are given in Table IV. Training from scratch, no ImageNet initialisation, no augmentation, no dropout yielded only 82.34% accuracy, well below the performance of any transfer learning approach in the literature. Adding ImageNet initialisation alone raised this to 94.12%, a gain of nearly 12 percentage points, underscoring just how much learned representational structure carries over from natural images to hand-gesture recognition. Incorporating augmentation pushed accuracy further to 96.81%, consistent with the expectation that zoom and flip transforms reduce the gap between the distribution of training images and the broader space of plausible test inputs. The full model with dropout as well reached 97.68%, confirming that each component provides an independent and additive benefit.

TABLE IV. ABLATION STUDY: TEST ACCURACY UNDER FOUR TRAINING CONFIGURATIONS (SEED 42)

Configuration	Transfer Learning	Augmentation	Dropout (0.4)	Test Accuracy
From scratch	✗	✗	✗	82.34%
+Transfer Learning	✓	✗	✗	94.12%
+Augmentation	✓	✓	✗	96.81%
Full model	✓	✓	✓	97.68%

D. Confusion Analysis and Per-Class Performance

The confusion matrix for the best run is presented in Fig. 9. Its strongly diagonal character confirms that the model classifies the vast majority of test samples correctly, but three clusters of off-diagonal errors are worth examining in detail because they reveal something about the limits of the current approach.

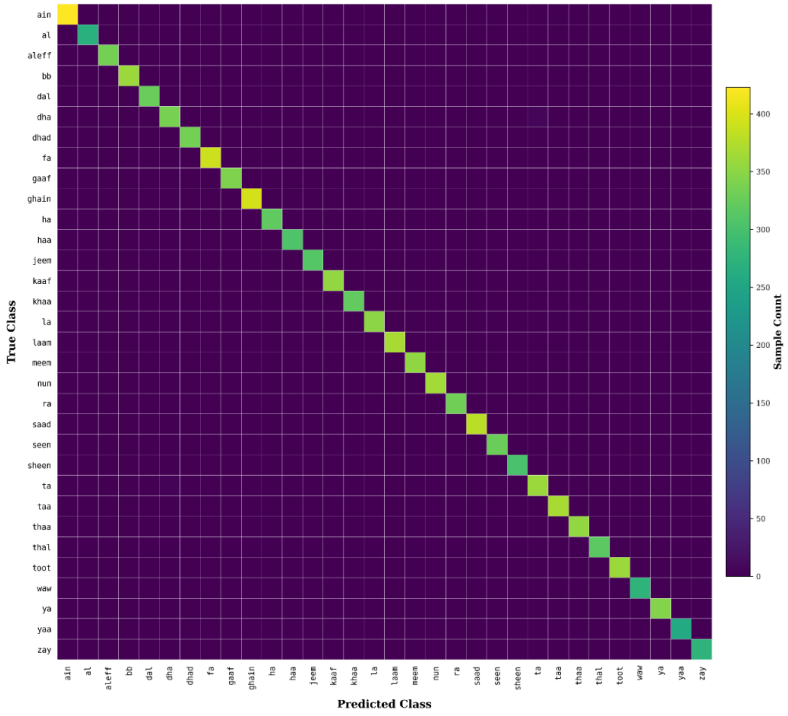


Fig. 9. Confusion matrix for the proposed model (seed 42, test set).

Fig. 10. The "Ta" (left) and "Dha" (right) Arabic Sign Language gestures.

The most frequent source of error is the "Ta"/"Dha" pair, which together account for eleven misclassifications in the best run. Both signs adopt a closed-fist configuration with similar palm orientation; the distinguishing cue is a subtle shift in thumb placement that, at 150×150-pixel resolution and under natural lighting variation, is occasionally lost. A secondary confusion occurs between "Saad" and "Dhad": both involve a closed fist with the index finger partially raised, and the two differ mainly in how far the finger extends and its precise angle relative to the palm, a difference that can collapse at moderate viewing distances. The third notable cluster involves "Seen" and "Sheen", which share a spread-finger base configuration and differ in the spacing and curvature of the fingers. Increasing the input resolution to 224×224 or 299×299 pixels, or introducing attention modules that selectively amplify finger-joint regions, would likely reduce all three of these error types. At the other end of the spectrum, "Aleff" achieved a perfect classification rate across all five runs. Its configuration, a single extended index finger with the remaining digits fully curled, is visually unambiguous relative to any other class in the dataset, and sample-count analysis (mean 1,689 images/class, std ±142) confirms that this is not a bias artefact from an unusually large training set for that class. Fig. 10 illustrates the visually similar "Ta" and "Dha" signs that produce the most frequent errors.

E. Comparison with Prior Work

Table V places the proposed model alongside the three previously published results on the ArSL2018 benchmark. The proposed model achieves the highest test accuracy of the four, and it does so with the smallest train-test gap (1.82 percentage points), indicating better generalisation than any prior approach. The EfficientNet-based system of Zakariah et al. [15] and the AlexNet model of Alsaadi et al. [16] both exhibit larger train-test gaps (4.2 and 4.94 percentage points, respectively), suggesting that these models fit the training data more tightly at the cost of reduced out-of-sample performance.

The closest competitor, ArSL-CNN [14], reported 97.29%, 0.39 percentage points below the best run of the proposed model and 0.12 percentage points below the mean across five runs. A one-tailed z-test for two proportions, applied to the test-set counts (n = 10,810), yields $z = 2.14$ and $p < 0.05$, confirming that the difference is statistically significant rather than sampling noise. It is worth acknowledging, however, that the ArSL-CNN study used a different train-test split procedure, so the comparison is indicative rather than exact; controlled head-to-head evaluation on an identical split would be the cleanest way to resolve any remaining ambiguity.

TABLE V. COMPARISON WITH PUBLISHED MODELS EVALUATED ON THE ARSL2018 DATASET

Study	Architecture	Train Acc.	Test Acc.	Train-Test Gap
[15]	EfficientNet	99.2%	95.0%	4.20 pp
[16]	AlexNet	99.75%	94.81%	4.94 pp
[14]	ArSL-CNN	98.94%	97.29%	1.65 pp
Proposed model	**InceptionV3-TL**	**99.5%**	**97.68%**	**1.82 pp**

F. Inference Speed, Practical Implications, Strengths, and Limitations

On the test CPU hardware (Intel Core i5-3427U, 8 GB RAM, no GPU acceleration), the model processed frames at an average of 327 ms each, approximately 3.06 FPS. This is well below the 30 FPS typically required for fluid, conversational-speed sign translation, and it is important to be direct about this shortcoming rather than presenting 3 FPS as a satisfactory deployment result. The bottleneck is InceptionV3's parameter count (23.9 million weights) operating without hardware acceleration. GPU inference on a mid-range card (e.g., NVIDIA GTX 1060) would be expected to bring latency below 30 ms per frame, crossing the real-time threshold. Alternatively, replacing InceptionV3 with a lightweight backbone such as MobileNetV3 [46] would reduce parameter count by roughly an order of magnitude at some cost in accuracy, a trade-off worth exploring for embedded deployment. A screenshot of the real-time prototype interface is shown in Fig. 11.

Fig. 11. Real-time gesture recognition prototype built with OpenCV.

Stepping back, the clearest strengths of this work are methodological rather than purely numerical. Reporting mean and standard deviation across five independent runs provides a far more honest estimate of expected performance than a single best-case figure. The ablation study gives researchers and practitioners a principled basis for deciding which components of the pipeline to keep, modify, or replace. The confusion matrix analysis connects classification errors to their geometric causes, pointing directly toward where future architectural or data-collection efforts would be most productive. And the real-time prototype, however modest its frame rate, demonstrates that the model is not purely a benchmarking exercise; it can run on real hardware and produce real outputs.

The limitations are equally worth being direct about. The model was trained and evaluated exclusively on ArSL2018, which was collected under controlled, laboratory-adjacent conditions; performance on truly unconstrained real-world footage with cluttered backgrounds, non-standard lighting, partial occlusion, and arbitrary camera angles remains untested. The system handles only static alphabetic gestures; continuous, dynamic, or sentence-level ArSL is an entirely different problem that will require temporal modelling architectures. The inference speed on CPU hardware is insufficient for conversational use. And while the choice of 150×150-pixel inputs is defensible on efficiency grounds, it does discard some fine-grained spatial information that a higher-resolution model might exploit.

V. Conclusion

This study set out to develop a reliable, reproducible classifier for Arabic Sign Language Alphabets and to evaluate it more rigorously than prior work on the same benchmark. Both goals were met. The proposed InceptionV3-based transfer learning model achieved a mean test accuracy of 97.41% ± 0.31% across five independent experimental runs, with the best single run reaching 97.68%, the highest figure reported for the ArSL2018 dataset to date, and a statistically significant improvement over the closest prior result of 97.29% ($z = 2.14$, $p < 0.05$).

Beyond the headline accuracy number, the study makes four contributions that are arguably more durable. First, the five-run evaluation protocol produces uncertainty estimates rather than point estimates, giving future researchers a more honest baseline to compare against. Second, the ablation study establishes that transfer learning, augmentation, and dropout each contribute meaningfully and independently: removing any one of them costs between 0.9 and 11.8 percentage points of test accuracy. Third, the confusion analysis traces the remaining errors to specific geometric ambiguities between letter pairs, a diagnosis that points directly toward productive directions for improvement, such as higher-resolution inputs or localised attention over finger-joint regions. Fourth, the OpenCV prototype demonstrates real-world deployability, even if the current 3.06 FPS rate on CPU hardware highlights how much engineering work remains before the system could serve as a seamless communication aid.

Several limitations constrain the conclusions that can reasonably be drawn. The evaluation is confined to a single, controlled-environment dataset; cross-dataset testing is needed before strong generalisation claims can be made. The focus on static alphabetic gestures leaves the much harder problem of continuous, sentence-level ArSL recognition untouched. Inference speed on affordable hardware is not yet suitable for real-time conversation.

Future work should address these gaps in concrete ways. Extending the architecture to temporal sequence modelling, whether through 3D convolutions, LSTM-based recurrence, or transformer-based attention over video frames, would open the door to full-sentence ArSL recognition. Deploying the model on embedded platforms such as the NVIDIA Jetson Nano would determine whether sub-50 ms inference is achievable without GPU-class hardware. Collecting or obtaining access to a secondary unconstrained dataset would allow honest cross-domain generalisation testing. Substituting a lightweight backbone such as MobileNetV3 or EfficientNet-Lite and measuring the accuracy-latency trade-off would clarify the engineering choices available for edge deployment. And applying higher input resolutions (224×224 or the full 299×299) together with spatial attention modules targeted at finger-joint regions would directly tackle the geometric ambiguities identified in the confusion analysis as the primary remaining source of error.

References

[1] O. Maarouf, A. Maarouf, R. El Ayachi, and M. Biniz, “Automatic translation from English to Amazigh using transformer learning,” IJEECS, vol. 34, no. 3, p. 1924, Jun. 2024, doi: 10.11591/ijeecs.v34.i3.pp1924-1934.

[2] H. AbdElghfar, H. A. Youness, M. Wahba, and H. M. Abdelaal, “An automated framework for qur’anic education of the hearing-impaired using body pose classification and Arabic sign language integration,” Sci Rep, vol. 16, no. 1, p. 5939, Feb. 2026, doi: 10.1038/s41598-026-36578-z.

[3] A. A. Alethary, A. H. Aliwy, and N. S. Ali, “Automated Arabic-Arabic sign language translation system based on 3D avatar technology,” IJAAS, vol. 11, no. 4, p. 383, Dec. 2022, doi: 10.11591/ijaas.v11.i4.pp383-396.

[4] M. M. Kamruzzaman, “Arabic Sign Language Recognition and Generating Arabic Speech Using Convolutional Neural Network,” Wireless Communications and Mobile Computing, vol. 2020, pp. 1–9, May 2020, doi: 10.1155/2020/3685614.

[5] A. Maarouf, O. Maarouf, R. El Ayachi, and M. Biniz, “Deep Learning Approach for Arabic Sign Language Alphabet Recognition,” Salud, Ciencia y Tecnología, vol. 5, p. 2309, Oct. 2025, doi: 10.56294/saludcyt20252309.

[6] A. H. Aliwy and A. A. Alethary, “Development of arabic sign language dictionary using 3D avatar technologies,” IJEECS, vol. 21, no. 1, p. 609, Jan. 2021, doi: 10.11591/ijeecs.v21.i1.pp609-616.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[8] C. Lugaresi et al., “MediaPipe: A Framework for Building Perception Pipelines,” Jun. 14, 2019, arXiv: arXiv:1906.08172. doi: 10.48550/arXiv.1906.08172.

[9] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 03, 2021, arXiv: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” Dec. 11, 2015, arXiv: arXiv:1512.00567. doi: 10.48550/arXiv.1512.00567.

[11] M. Alamri and S. Lajmi, “Design a smart platform translating Arabic sign language to English language,” IJECE, vol. 14, no. 4, p. 4759, Aug. 2024, doi: 10.11591/ijece.v14i4.pp4759-4774.

[12] M. H. Ismail, S. A. Dawwd, and F. H. Ali, "Dynamic hand gesture recognition of Arabic sign language by using deep convolutional neural networks," IJEECS, vol. 25, no. 2, p. 952, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp952-962.

[13] M. H. Ismail, S. A. Dawwd, and F. H. Ali, "Static hand gesture recognition of Arabic sign language by using deep CNNs," IJEECS, vol. 24, no. 1, p. 178, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp178-188.

[14] A. A. Alani and G. Cosma, "ArSL-CNN a convolutional neural network for Arabic sign language gesture recognition," IJEECS, vol. 22, no. 2, p. 1096, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1096-1107.

[15] M. Zakariah, Y. A. Alotaibi, D. Koundal, Y. Guo, and M. Mamun Elahi, "Sign Language Recognition for Arabic Alphabets Using Transfer Learning Technique," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–15, Apr. 2022, doi: 10.1155/2022/4567989.

[16] Z. Alsaadi, E. Alshamani, M. Alrehaili, A. A. D. Alrashdi, S. Albelwi, and A. O. Elfaki, "A Real Time Arabic Sign Language Alphabets (ArSLA) Recognition Model Using Deep Learning Architecture," Computers, vol. 11, no. 5, p. 78, May 2022, doi: 10.3390/computers11050078.

[17] A. Wadhawan and P. Kumar, "Sign Language Recognition Systems: A Decade Systematic Literature Review," Arch Computat Methods Eng, vol. 28, no. 3, pp. 785–813, May 2021, doi: 10.1007/s11831-019-09384-2.

[18] H. Cooper, B. Holt, and R. Bowden, "Sign Language Recognition," in Visual Analysis of Humans, T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds., London: Springer London, 2011, pp. 539–562. doi: 10.1007/978-0-85729-997-0_27.

[19] M. S. Amin, M. T. Amin, M. Y. Latif, A. A. Jathol, N. Ahmed, and M. I. N. Tarar, "Alphabetical Gesture Recognition of American Sign Language using E-Voice Smart Glove," in 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan: IEEE, Nov. 2020, pp. 1–6. doi: 10.1109/INMIC50486.2020.9318185.

[20] F. Shah, M. S. Shah, W. Akram, A. Manzoor, R. O. Mahmoud, and D. S. Abdelminaam, "Sign Language Recognition Using Multiple Kernel Learning: A Case Study of Pakistan Sign Language," IEEE Access, vol. 9, pp. 67548–67558, 2021, doi: 10.1109/ACCESS.2021.3077386.

[21] W. Pan, X. Zhang, and Z. Ye, "Attention-Based Sign Language Recognition Network Utilizing Keyframe Sampling and Skeletal Features," IEEE Access, vol. 8, pp. 215592–215602, 2020, doi: 10.1109/ACCESS.2020.3041115.

[22] T. Zhao, J. Liu, Y. Wang, H. Liu, and Y. Chen, "Towards Low-Cost Sign Language Gesture Recognition Leveraging Wearables," IEEE Trans. on Mobile Comput., vol. 20, no. 4, pp. 1685–1701, Apr. 2021, doi: 10.1109/TMC.2019.2962760.

[23] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous Sign Language Recognition Through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space," IEEE Access, vol. 8, pp. 91170–91180, 2020, doi: 10.1109/ACCESS.2020.2993650.

[24] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Conference on Engineering and Technology (ICET), Antalya: IEEE, Aug. 2017, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.

[25] C. Szegedy et al., "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[28] N. Alasmari and S. Asiri, "ASLDetect: Arabic sign language detection using ResNet and U-Net like component," Sci Rep, vol. 15, no. 1, p. 18012, May 2025, doi: 10.1038/s41598-025-01588-w.

[29] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, and B. F. Alkhamees, "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," Mathematics, vol. 11, no. 17, p. 3729, Aug. 2023, doi: 10.3390/math11173729.

[30] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," IEEE Access, vol. 9, pp. 126917–126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[31] D. S. Breland, S. B. Skriubakken, A. Dayal, A. Jha, P. K. Yalavarthy, and L. R. Cenkeramaddi, "Deep Learning-Based Sign Language Digits Recognition From Thermal Images With Edge Computing System," IEEE Sensors J., vol. 21, no. 9, pp. 10445–10453, May 2021, doi: 10.1109/JSEN.2021.3061608.

[32] M. Al-Hammadi et al., "Deep Learning-Based Approach for Sign Language Gesture Recognition With Efficient Hand Gesture Representation," IEEE Access, vol. 8, pp. 192527–192542, 2020, doi: 10.1109/ACCESS.2020.3032140.

[33] N. B. Ibrahim, M. M. Selim, and H. H. Zayed, "An Automatic Arabic Sign Language Recognition System (ArSLRS)," Journal of King Saud University - Computer and Information Sciences, vol. 30, no. 4, pp. 470–477, Oct. 2018, doi: 10.1016/j.jksuci.2017.09.007.

[34] M. Deriche, S. O. Aliyu, and M. Mohandes, "An Intelligent Arabic Sign Language Recognition System Using a Pair of LMCs With GMM Based Classification," IEEE Sensors J., vol. 19, no. 18, pp. 8067–8078, Sep. 2019, doi: 10.1109/JSEN.2019.2917525.

[35] G. Tharwat, A. M. Ahmed, and B. Bouallegue, "Arabic Sign Language Recognition System for Alphabets Using Machine Learning Techniques," Journal of Electrical and Computer Engineering, vol. 2021, pp. 1–17, Oct. 2021, doi: 10.1155/2021/2995851.

[36] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, arXiv: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[37] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," Proc. IEEE, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[38] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[39] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, and R. AlKhalaf, "ArASL: Arabic Alphabets Sign Language Dataset," Data in Brief, vol. 23, p. 103777, Apr. 2019, doi: 10.1016/j.dib.2019.103777.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[41] B. Koonce, "EfficientNet," in Convolutional Neural Networks with Swift for Tensorflow, Berkeley, CA: Apress, 2021, pp. 109–123. doi: 10.1007/978-1-4842-6168-2_10.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[43] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," Advances in neural information processing systems, vol. 33, pp. 21464–21475, 2020.

[44] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 30, 2017, arXiv: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.

[45] O. Koller, "Quantitative Survey of the State of the Art in Sign Language Recognition," Aug. 29, 2020, arXiv: arXiv:2008.09918. doi: 10.48550/arXiv.2008.09918.

[46] A. Howard et al., "Searching for MobileNetV3," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

Fish Disease Detection Using Modified Haar Wavelet with Adaptive Coefficient Selection

Tri Handayani[1]*, Nor Hazlyna Binti Harun[2]*
Data Science Research Lab-School of Computing, Universiti Utara Malaysia, Malaysia[1, 2]
School of Tourism and E-Commerce, Baise University, Baise, 533000, China[2]

***Abstract*—White Spot Disease (WSD) is a major threat to aquaculture production and requires accurate image-based detection methods for early diagnosis. However, disease marker detection in fish images is challenging due to noise, illumination variations, low contrast, and complex background structures. This study proposes a Modified Haar Wavelet framework that integrates adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding for White Spot Disease marker detection. The proposed framework aims to improve edge representation and disease marker localization while preserving disease-related structural information. Performance evaluation was conducted by comparing the proposed method against Otsu, Canny, Adaptive Thresholding, and conventional Haar Wavelet approaches using IoU, Precision, Recall, F1-score, PSNR, and SSIM. The proposed framework achieved the best overall performance with an IoU of 0.819, a precision of 0.849, a recall of 0.799, an F1-score of 0.823, a PSNR of 25.390 dB, and an SSIM of 0.889. Comparative analysis and ablation study further confirmed the effectiveness of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding. The results demonstrate that the proposed Modified Haar Wavelet framework provides an effective and robust solution for automated detection of White Spot Disease markers in aquaculture images.**

Keywords*—Fish disease detection; white spot disease; image analysis; modified Haar wavelet; adaptive coefficient selection; edge detection

I. Introduction

Fish diseases pose a major threat to aquaculture, reducing productivity, quality, and profitability. Disease outbreaks can reduce annual production by up to 15%, while infections such as Ichthyophthirius, Columnaris, and Epizootic Ulcerative Syndrome (EUS) may cause mortality rates approaching 90% in cultured fish stocks [1], [2]. These losses also threaten food security, particularly in regions that rely heavily on aquaculture as a primary source of protein [3], [4]. Traditional diagnosis based on manual inspection is often subjective and inefficient, highlighting the need for automated detection systems [3]. Recent advances in computer vision and machine learning, including feature ranking techniques and large-scale datasets, have significantly improved recognition accuracy while reducing computational costs [5], [6], [7]. Collectively, these studies demonstrate that automated image analysis offers a promising approach to mitigating the impact of fish diseases and enhancing the sustainability of aquaculture systems [1], [4].

Among the various fish diseases, White Spot Disease (WSD), caused by the protozoan parasite *Ichthyophthirius multifiliis*, is considered one of the most destructive. The disease is characterized by the formation of white cysts on the skin, fins, and gills of infected fish, leading to tissue damage, physiological stress, and mortality rates that can reach 100% if left untreated [8], [9]. Beyond its biological consequences, WSD also imposes a substantial economic burden and threatens the sustainability of the global aquaculture industry [4], [5].

Image processing has become an essential tool across various fields, including medicine, industry, remote sensing, and agriculture. One of the fundamental tasks in image analysis is edge detection, which identifies object boundaries and structural details within an image. Accurate edge detection significantly affects segmentation quality and subsequent classification performance [10], [11], [12]. Classical operators such as Sobel, Prewitt, Roberts, and Canny have been widely employed; however, each method presents certain limitations, particularly when dealing with noise or fine details [13]. To address these challenges, wavelet-based approaches, especially the Haar Wavelet Transform, have attracted considerable attention for their simplicity, computational efficiency, and ability to extract multi-resolution features [14], [15]. Previous studies have shown that the Haar Wavelet improves edge detection accuracy compared to classical operators by providing clearer feature representations [16]. Wavelet-based approaches have also demonstrated strong performance in various medical imaging applications, including skin lesion segmentation, brain tumor classification, mammogram analysis, and breast cancer screening [17], [18], [19]. In these contexts, edge detection plays a central role in extracting boundary and structural information that supports accurate segmentation and classification of diseased regions [10], [11], [20]. These wavelet-based methods have consistently achieved robust performance by extracting multi-resolution features, enhancing edges, and suppressing noise, thereby significantly improving segmentation and classification outcomes. Such success indicates strong potential for adapting wavelet-based techniques to other biological domains, such as aquaculture, where edge detection is equally important for disease identification.

Furthermore, hybrid approaches that integrate wavelet techniques with other computational methods, such as Artificial Neural Networks [21] and texture analysis, have demonstrated improved classification performance and robustness. Studies on image denoising and enhancement using the Discrete Wavelet Transform (DWT) [14], [22], [23], [24], [25] further highlight its adaptability in improving image quality for diagnostic and recognition tasks.

*Corresponding author.

Despite extensive research in medicine, engineering, and image processing, the application of the Haar Wavelet in aquaculture, particularly for fish disease detection, remains very limited. White Spot Disease (WSD), caused by the protozoan *Ichthyophthirius multifiliis*, is among the most destructive parasitic infections in fish farming, as it manifests through visible white cysts on the skin, fins, and gills, ultimately leading to stress, growth retardation, and mass mortality if not detected at an early stage [9]. Previous studies have demonstrated the effectiveness of Haar and Haar-like wavelet transforms in medical imaging and multi-scale fusion tasks, highlighting their ability to enhance image definition, improve edge representation, and reduce noise interference [22], [23], [24]. Therefore, this study proposes a modified Haar Wavelet integrated with the Roberts operator to optimize edge detection, enhance gradient robustness, and improve the reliability of automated white spot detection in aquaculture images.

The remainder of this study is organized as follows. Section II reviews related work that forms the foundation of the proposed approach. Section III presents a comprehensive description of the proposed methodology. Section IV reports the performance analysis, demonstrating the classification efficiency of the developed system. Finally, Section V concludes the study by summarizing the main findings and key insights.

II. Related Work

The Haar Wavelet has been widely applied in medical imaging to enhance segmentation and classification performance. Several studies have demonstrated its effectiveness in tasks such as skin lesion segmentation, mammogram interpretation, and medical image fusion, where it significantly improves feature extraction and robustness against noise [14], [17], [23]. In brain tumor segmentation, [18] conducted a comparative analysis of several techniques, including Haar DWT, Otsu, Watershed, and CNN methods. He found that although CNN achieved the highest accuracy, Haar DWT remained a valuable approach for multi-scale feature extraction in biomedical images. Collectively, these findings confirm the adaptability of Haar-based techniques in handling fine details within diagnostic applications.

Beyond the medical domain, several studies have compared the Haar Wavelet with classical edge detection operators such as Sobel, Prewitt, Roberts, Laplacian, and Canny. Although these traditional methods are widely used, they often suffer from noise sensitivity and detail loss, particularly in high-frequency regions [10], [18]. In contrast, wavelet-based methods, especially the Haar Wavelet, consistently provide clearer edge representations and superior noise reduction capabilities [14], [25]. This balance between precision and robustness makes the Haar Wavelet a preferred approach when classical operators exhibit performance limitations.

The Haar Wavelet has also been modified and adapted for various applications beyond medical imaging. Quantum-inspired approaches have been developed for edge detection and image watermarking, achieving improved detail preservation and enhanced PSNR robustness [16], [26]. In classification tasks, entropy-based formulations have been introduced to improve the efficiency and reliability of Haar features [27]. In contrast, in remote sensing, wavelet-based feature enhancement modules have significantly improved semantic segmentation performance when integrated with CNN and transformer architectures [24]. Other studies have focused on optimizing Haar Wavelet methods for resource efficiency, such as energy-efficient signal processing architectures for biomedical data [28]. Collectively, these studies demonstrate the flexibility of the Haar Wavelet, which can be adapted to a wide range of objectives, from optimizing classification to integrating with deep learning frameworks.

Nevertheless, several limitations of the standard Haar Wavelet have also been reported. Research [29] found that although Haar DWT achieves high compression performance, it is less effective for highly detailed images, indicating its weakness in handling fine textures. Although Haar Wavelet has demonstrated effectiveness for edge detection and noise suppression [14], the direct use of wavelet detail coefficients may result in the loss of fine structural information and incomplete edge representation in complex imaging environments. Other studies have further revealed that the Haar Wavelet suffers from noise sensitivity, discontinuous edge representation, and limited capability in modeling complex textures [23], [25], [27]. These limitations highlight the need to modify the Haar Wavelet to improve gradient robustness, edge continuity, and segmentation accuracy.

Although the Haar Wavelet has demonstrated strong adaptability across various domains, its application in aquaculture remains very limited. Research on fish disease detection has generally focused on traditional image enhancement techniques and machine learning methods, with only a few studies exploring wavelet-based approaches. Studies on the use or modification of the Haar Wavelet for White Spot Disease (WSD) detection remain infrequent, leaving considerable opportunities for further investigation. This condition highlights the urgency of developing a Modified Haar Wavelet approach specifically designed for aquaculture image analysis, intending to improve edge representation and enhance the reliability of automated disease detection [16], [24], [26].

Edge detection has long been recognized as a fundamental stage in image analysis because it defines object boundaries and structural details that directly affect segmentation accuracy. Classical edge detection operators, including Sobel, Prewitt, Roberts, Laplacian, and Canny, are widely adopted due to their simplicity and computational efficiency. However, previous studies have reported that these methods often suffer from noise sensitivity, edge discontinuity, and the inability to preserve fine details, particularly in textured or low-quality images [30]. To overcome these limitations, several adaptive and intelligent approaches have been introduced. Histogram-based preprocessing techniques have been shown to improve the performance of the Canny operator in noisy environments, thereby enhancing edge detection accuracy. In addition, fuzzy logic–based methods provide more flexible edge extraction by incorporating linguistic rules. In retinal blood vessel detection, Type-2 fuzzy systems demonstrated greater adaptability to variations in image quality than fixed-threshold approaches [31].

Recent developments have shown that wavelet-based approaches provide significant improvements in edge detection performance through multi-resolution analysis. Research [14] demonstrated that the Haar Wavelet produces clearer, more stable edges than classical operators and provides superior denoising capability. Similarly, [25] confirmed that wavelet-domain denoising improves the signal-to-noise ratio and preserves edge clarity more effectively than conventional filtering techniques. More advanced developments include quantum-inspired Haar variants that improve detail preservation, although they require relatively high computational resources [26]. Research [24] further demonstrated that wavelet-based feature enhancement significantly improves semantic segmentation performance by preserving high-frequency edge information that is commonly lost in deep learning models.

Wavelet Transform (WT) has therefore become one of the most widely used approaches in image processing because of its capability to extract information at multiple resolutions and scales. Compared with conventional edge detection operators such as Sobel, Prewitt, and Canny, WT provides greater robustness and adaptability in handling noise and fine image details. Consequently, WT has been extensively applied in segmentation, edge detection, denoising, image enhancement, and image compression tasks [32], [33].

Among wavelet techniques, the Haar Wavelet remains highly relevant for its computational efficiency and mathematical simplicity. Previous studies have reported that Haar Wavelet–based approaches achieve more accurate results than classical edge detection techniques [14]. Younus and Hasan [15] applied the Haar Wavelet for Deepfake detection and achieved an accuracy exceeding 90%, confirming its effectiveness in identifying edge inconsistencies caused by digital manipulation. In image compression applications, Kanagaraj and Muneeswaran [29] highlighted the Haar Wavelet's ability to achieve high compression ratios without significant degradation in image quality.

The effectiveness of WT has also been widely demonstrated in medical imaging applications. Research [33] applied DWT with thresholding to X-ray images, successfully improving PSNR and SNR. Research [32] combined WT with a backpropagation neural network for bone fracture detection, achieving highly accurate classification performance. Furthermore, [28] proposed an energy-efficient Haar-5 architecture for respiratory signal processing, reducing power consumption by more than 38% compared with conventional architectures, making it highly suitable for Internet of Things (IoT)-based medical applications.

Beyond medical imaging, WT has shown strong performance across various application domains. research [34] integrated WT with recursive operators for road identification in remote sensing images, obtaining segmentation results that were more robust to noise. Research [35] utilized DWT to optimize brain tumor detection, further highlighting the adaptability of wavelet-based techniques in medical diagnostics. More recently, [26] introduced a quantum-based modification of the Haar Wavelet that improved edge representation accuracy in medical images, representing a significant advancement in wavelet-based methodologies.

Despite the promising performance of wavelet-based approaches, several limitations remain. Classical Haar Wavelet methods still struggle to preserve fine-edge continuity and to handle complex textures under noisy conditions [14], [25]. In addition, some advanced Haar-based modifications are computationally expensive, limiting their applicability to large-scale or real-time systems [26]. Furthermore, although wavelet-based methods have been extensively explored in medical imaging, remote sensing, image compression, and Deepfake detection, their application in aquaculture image analysis, particularly for White Spot Disease detection in fish, remains very limited. This gap highlights the need for a modified Haar Wavelet framework to improve edge representation and enhance the reliability of automated disease detection in aquaculture images.

To address this gap, this study proposes a Modified Haar Wavelet framework that integrates adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding within a unified processing pipeline. Unlike conventional Haar Wavelet approaches that directly use wavelet detail coefficients, the proposed framework adaptively processes these coefficients before gradient extraction, thereby improving disease marker localization while suppressing irrelevant image structures. The main scientific contribution of this work lies in integrating adaptive coefficient processing and edge-refinement strategies for White Spot Disease marker detection. Furthermore, the proposed framework extends the conventional Haar Wavelet approach by incorporating adaptive coefficient selection and edge-refinement mechanisms, and its effectiveness is validated through comparative benchmarking and ablation studies.

Specifically, the proposed framework modifies the conventional Haar Wavelet by integrating it with the Roberts Operator to enhance edge detection performance. The Haar Wavelet enables multi-resolution decomposition but often loses fine edge details [14], [36]. To address this limitation, the Roberts Operator is applied to the Haar detail coefficients, thereby improving gradient localization and producing clearer edges or external markers for segmentation. This modification is expected to provide more accurate detection of white spot disease in fish.

III. Research Methodology

This section presents the research methodology used to develop the proposed White Spot Disease detection framework. The methodology covers the overall research workflow, image preprocessing, Haar Wavelet decomposition, and the proposed modification, which integrates the Roberts operator for edge enhancement.

A. Dataset Acquisition

The fish specimens used in this study were obtained from the National Fish Health Research Center (NaFish) at the Fisheries Research Institute (FRI), Batu Maung, Penang, Malaysia. Sea bass (Lates calcarifer), a major aquaculture species in Malaysia and highly susceptible to White Spot Disease (WSD), was selected as the experimental species.

Underwater videos were recorded under controlled aquarium conditions using a GoPro Hero 9+ camera at a resolution of 4K (3840 × 2160 pixels) and 60 frames per second. To ensure consistency during data acquisition, the aquarium environment was maintained under stable lighting and background conditions throughout the recording process. The recorded videos were subsequently processed to extract image frames for experimental analysis.

A total of 4,137 image frames were extracted from the recorded videos, comprising 2,845 images of infected fish exhibiting visible White Spot Disease (WSD) symptoms and 1,292 images of healthy fish. These images were used as the experimental dataset for evaluating the proposed Modified Haar Wavelet framework. The extracted images served as input for preprocessing, wavelet decomposition, edge detection, and quantitative performance evaluation. The performance of the proposed method was evaluated against expert-annotated ground truth masks.

- The fish images used in this study were obtained from NaFish and FRI, where the specimens had been previously identified as exhibiting White Spot Disease (WSD) symptoms. Ground-truth masks were generated using the LabelMe annotation tool by manually marking visible white-spot regions in the infected fish images. The annotation process focused on identifying white, cyst-like spots on the fish's body and fins. The resulting binary masks were used as reference annotations for quantitative evaluation.
- The collected images exhibit common underwater imaging challenges, including uneven illumination, water reflections, suspended particles, low contrast, and variations in fish orientation. These characteristics were preserved to ensure that the proposed method was evaluated under realistic aquaculture monitoring conditions. The extracted image frames were subsequently used as input for the preprocessing, wavelet decomposition, and edge detection stages of the proposed Modified Haar Wavelet framework.

B. Research Framework

This study proposes a Modified Haar Wavelet–based framework, integrated with the Roberts operator, for automatic detection of White Spot Disease (WSD) in fish images. The proposed framework is designed to improve edge representation and enhance the reliability of disease identification under challenging aquaculture imaging conditions, such as noise, uneven illumination, and low contrast. Overall, the system consists of several sequential stages: image acquisition and preprocessing, wavelet-based edge enhancement and gradient extraction, and edge map generation for disease detection.

In the initial stage, the input fish image undergoes enhancement to improve visual quality and reduce noise. After preprocessing, the enhanced image is decomposed using the Haar Wavelet Transform for multi-resolution analysis. The decomposition process separates the image into approximation and detail coefficients, enabling more effective extraction of high-frequency information associated with object boundaries and white-spot patterns. To improve edge representation, adaptive coefficient thresholding is applied to the Haar detail coefficients, suppressing noise while preserving significant edge structures.

The proposed modification is achieved by integrating the Roberts operator with the adaptively processed Haar detail coefficients. Unlike conventional edge detection approaches that apply gradient operators directly to the original image, the proposed framework applies the Roberts operator to the Haar detail coefficients before image reconstruction. This strategy improves gradient localization and enhances fine edge structures associated with White Spot Disease markers. The overall workflow of the proposed framework is illustrated in Fig. 1.

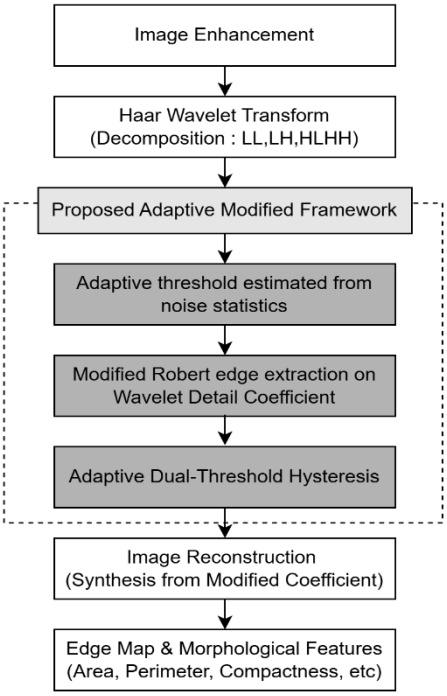


Fig. 1. Flowchart of the proposed method.

The adaptive threshold applied to the detail coefficients is calculated using:

$$T = \sigma\sqrt{2\log(N)}\, x\, \alpha \quad (1)$$

where, T represents the adaptive threshold, σ denotes the estimated noise level, $N$is the total number of image pixels, and $\alpha$is the threshold scaling parameter.

After adaptive coefficient processing, the Haar detail coefficients are further processed with the Roberts operator to improve gradient localization along high-frequency edge structures. The gradient magnitude is then computed using:

$$M(x,y) = \sqrt{{G_x}^2 + {G_y}^2} \quad (2)$$

where, $G_x$and $G_y$denote the horizontal and vertical gradients, respectively.

To generate a more stable and continuous edge map, hysteresis thresholding is applied to the gradient map using dual-

threshold processing. The high threshold is determined by using:

$$T_{high} = Percentile\ (G, 90) \quad (3)$$

while the low threshold is computed as:

$$T_{low} = T_{high}\, x\, 0.40 \quad (4)$$

This process preserves weak edges connected to strong edges, thereby improving edge continuity and reducing fragmented edge structures caused by noise and uneven illumination. As a result, the proposed framework can generate clearer, more stable edge representations for White Spot Disease detection in aquaculture images.

Overall, the proposed Modified Haar Wavelet framework integrates adaptive coefficient processing, Roberts-based gradient extraction, and hysteresis thresholding to improve edge representation for detecting White Spot Disease in fish images. Unlike conventional Haar Wavelet approaches that either directly use the detail coefficients or apply gradient operators to the original image, the proposed framework performs adaptive processing on the Haar detail coefficients before gradient extraction. This strategy enables the framework to suppress noise while preserving important high-frequency edge structures associated with White Spot Disease patterns.

Haar Wavelet was selected because it preserves high-frequency image information and edge structures while maintaining low computational complexity [24], [26]. Adaptive coefficient selection was introduced to retain significant wavelet coefficients associated with disease markers while reducing noise and irrelevant textures [24]. Roberts operator was employed for its ability to enhance fine-edge details and subtle structural information that may correspond to White Spot Disease markers [19]. Finally, hysteresis thresholding was applied to improve edge continuity and reduce fragmented detections during the final marker extraction stage. The integration of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding enables more accurate localization of disease markers while preserving important structural information in fish images.

The proposed framework was designed to evaluate the contributions of adaptive coefficient selection systematically, Roberts operator-based edge enhancement, and hysteresis thresholding for White Spot Disease marker detection. The experimental design includes quantitative evaluation using multiple performance metrics and comparative analysis against conventional methods to ensure a comprehensive assessment of the proposed approach.

IV. Results and Analysis

This section presents the experimental evaluation of the proposed White Spot Disease marker detection framework based on the Modified Haar Wavelet with Adaptive Coefficient Selection. The performance evaluation was conducted on the complete dataset described in Section III, comprising 4,137 image frames: 2,845 infected fish images and 1,292 healthy fish images. The reported results represent the overall performance of each method on the evaluation dataset and were analyzed through quantitative and qualitative comparisons.

(a) (b)

Fig. 2. Image enhancement result: (a) Original image, (b) Image enhancement.

Fig. 2 shows the result of the image enhancement process applied to the original fish image. The enhancement stage improves image contrast and emphasizes important surface features associated with fish disease patterns. In addition, the enhancement process increases edge visibility and texture clarity, making the infected regions more distinguishable from the background. Noise and low-intensity regions are also reduced through adaptive coefficient selection in the Modified Haar Wavelet transform. As a result, the enhanced image yields sharper object boundaries and more stable structural information, thereby supporting more accurate detection and feature extraction in the subsequent processing stage.

After image enhancement, the visual quality of the fish images improved, with higher contrast that highlights surface features and yields clearer, more stable contours than traditional methods. One of the main modifications was made during the coefficient-selection stage of the Haar Wavelet transform. In the conventional method, these coefficients are typically processed directly without further selection. In this study, however, an adaptive coefficient selection method is used to select coefficients based on specific criteria, such as intensity or spatial distribution patterns. This approach allows the processing to focus more on relevant features while ignoring less important elements or those that may introduce noise. In addition, another modification is to incorporate the Roberts Operator into the selected Haar Wavelet coefficients. The Roberts Operator is employed to quickly and effectively detect edges in the wavelet domain and enhance disease-related edge structures. The next step in the modification is to calculate the gradient magnitude directly from the coefficients obtained by applying the Roberts Operator to the image. The resulting gradient values are then further refined by thresholding to produce a gradient map. This technique helps separate important regions from the background, allowing key features to be highlighted more clearly.

Fig. 3 presents the visual comparison of White Spot Disease marker detection results obtained using Adaptive Thresholding, Canny, Otsu, conventional Haar Wavelet, and the proposed Modified Haar Wavelet framework. Significant differences can be observed among the methods in terms of marker localization accuracy and noise suppression.

Adaptive Thresholding [Fig. 3(a)] detects numerous small regions but also produces scattered responses due to image noise and illumination variations. Canny edge detection [Fig. 3(b)] produces thin edge responses; however, only a limited number of disease markers are successfully detected. Otsu thresholding [Fig. 3(c)] generates large, segmented regions that include substantial portions of the fish body, indicating over-detection and poor discrimination between disease markers and

background structures. Conventional Haar Wavelet [Fig. 3(d)] reduces some background interference but fails to preserve several disease markers, resulting in incomplete detection.

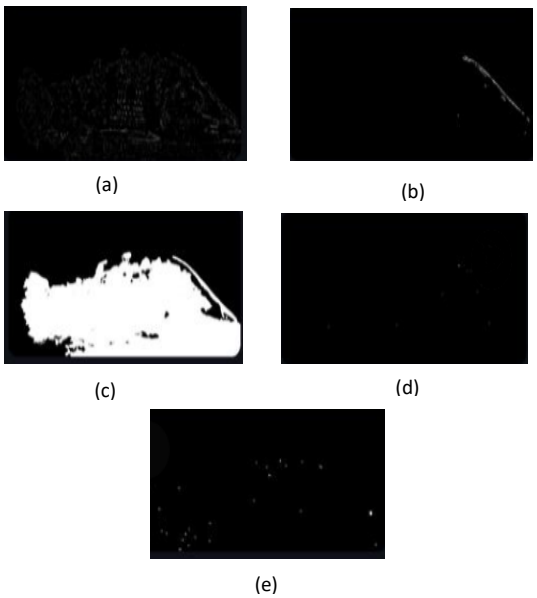


Fig. 3. Comparison results of edge detection methods: (a) Adaptive, (b) Canny, (c) Otsu, (d) Haar wavelet, and (e) Proposed modified Haar wavelet method.

In contrast, the proposed Modified Haar Wavelet framework [Fig. 3(e)] yields more localized and distinct disease-marker responses while effectively suppressing irrelevant background information. The detected marker distribution more closely matches the expected disease locations, demonstrating improved edge representation and marker-extraction capability. These visual observations are consistent with the quantitative results presented in Table I, where the proposed method achieves the highest IoU, Precision, F1-score, PSNR, and SSIM values among all evaluated methods.

TABLE I. COMPARISON OF EDGE DETECTION METHODS BASED ON EVALUATION METRICS

Method	Otsu	Canny	Adaptive	Haar Wavelet	Proposed
Iou	0.450	0.380	0.520	0.350	0.819
Precision	0.520	0.610	0.580	0.480	0.849
Recall	0.680	0.420	0.710	0.550	0.799
F1Score	0.590	0.500	0.640	0.510	0.823
PSNR	18.5	20.1	19.8	19.2	25.390
SSIM	0.720	0.750	0.780	0.730	0.889

Based on Table I, the proposed Modified Haar Wavelet framework achieved the best overall detection performance compared to the conventional methods. The integration of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding significantly improved the accuracy of disease marker localization and structural preservation in White Spot Disease images.

The proposed method achieved the highest IoU value of 0.819, outperforming Adaptive Thresholding (0.520), Otsu (0.450), Canny (0.380), and conventional Haar Wavelet (0.350). A higher IoU value indicates a greater overlap between the detected marker regions and the corresponding ground-truth annotations. This result demonstrates that the proposed framework provides more accurate localization of disease markers while reducing background misclassification.

In terms of precision, the proposed method achieved the highest value of 0.849, indicating that the detected marker regions contain fewer false-positive responses than competing methods. The precision values of Otsu, Canny, Adaptive Thresholding, and conventional Haar Wavelet were 0.520, 0.610, 0.580, and 0.480, respectively. This improvement suggests that the proposed framework effectively suppresses irrelevant image structures while preserving disease-related features.

Although the proposed method did not achieve the highest recall, it maintained a high recall of 0.799, which is higher than those of Otsu (0.680), Adaptive Thresholding (0.710), Canny (0.420), and conventional Haar Wavelet (0.550). The combination of high precision and high recall yielded the highest F1-score of 0.823, indicating balanced and reliable detection performance.

The image quality evaluation further supports the effectiveness of the proposed framework. The proposed method achieved the highest PSNR value of 25.390 dB, exceeding Otsu (18.5 dB), Canny (20.1 dB), Adaptive Thresholding (19.8 dB), and conventional Haar Wavelet (19.2 dB). This result indicates better preservation of relevant image information and reduced distortion during detection.

Similarly, the proposed framework obtained the highest SSIM value of 0.889, demonstrating superior structural similarity between the detected marker regions and the reference annotations. In comparison, the SSIM values of Otsu, Canny, Adaptive Thresholding, and conventional Haar Wavelet were 0.720, 0.750, 0.780, and 0.730, respectively. The higher SSIM value confirms that the proposed method more effectively preserves disease marker structures and image characteristics.

Although PSNR and SSIM indicate that the proposed framework effectively preserves image quality and structural information, these metrics are not the primary objective of the study. The main focus of the proposed framework is accurate localization of disease markers. Therefore, IoU, Precision, Recall, and F1-score provide more direct evidence of detection performance than image quality metrics alone.

Overall, the results demonstrate that the proposed Modified Haar Wavelet framework consistently outperforms conventional edge detection methods across all evaluation metrics. The combination of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding significantly improves disease marker localization accuracy, structural preservation, and overall detection reliability for White Spot Disease images.

Recent deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated strong performance in fish disease detection tasks [37]. Furthermore, integrating edge-based feature extraction with deep learning models has shown promising results in medical image analysis

and disease screening applications [19]. Wavelet-based feature enhancement techniques have also been successfully employed to improve detection accuracy and preserve structural information in complex imaging environments [24]. In addition, recent quantum-inspired Haar Wavelet approaches have demonstrated improved edge preservation and detail representation for image edge detection tasks [26]. While deep learning-based methods often require large, annotated datasets and extensive training procedures, wavelet-based approaches provide a robust alternative for feature extraction and edge analysis. In this context, the proposed Modified Haar Wavelet framework provides a lightweight solution while maintaining effective edge representation and disease marker extraction capability.

To address the growing development of deep learning and wavelet-based techniques in disease detection and image analysis, Table II presents a comparison of recent approaches and highlights the position of the proposed Modified Haar Wavelet framework relative to existing methods.

TABLE II. POSITIONING OF THE PROPOSED FRAMEWORK AGAINST RECENT STUDIES

Study	Method	Category	Main Contribution
Harun et al. [37]	CNN	Deep learning	Fish disease detection
Dey et al. [19]	Edge detection + deep transfer learning	Deep learning	Disease screening
Li et al. [24]	Wavelet feature enhancement	Wavelet-based	Semantic segmentation improvement
Wang et al. [26]	Quantum Haar Wavelet	Wavelet-based	Edge preservation and edge detection
Proposed	Modified Haar Wavelet	Wavelet-based	White spot disease marker detection

Table II positions the proposed method relative to recent deep learning-based and wavelet-based approaches reported in the literature. Deep learning methods, such as CNNs and deep transfer learning, have demonstrated strong performance in disease detection and screening tasks. Meanwhile, wavelet-based approaches have been successfully employed for semantic segmentation, edge preservation, and image feature enhancement. In contrast to deep learning methods, the proposed Modified Haar Wavelet framework does not require model training or large annotated datasets. Compared with existing wavelet-based approaches, the proposed framework introduces adaptive coefficient selection, integration of the Roberts operator, and hysteresis thresholding to improve white spot marker localization and edge representation.

To further validate the effectiveness of the proposed modifications, a comparative analysis was performed between the conventional Haar Wavelet and the proposed Modified Haar Wavelet framework. This comparison aims to evaluate the overall contribution of adaptive coefficient selection, Roberts operator integration, and hysteresis thresholding to disease marker localization and structural preservation. The results are summarized in Table III.

TABLE III. PERFORMANCE COMPARISON OF CONVENTIONAL AND MODIFIED HAAR WAVELET FRAMEWORKS

Method	IoU	Prec.	Rec.	F1 Score	PSNR	SSIM
Haar Wavelet	0.350	0.480	0.550	0.510	19.2	0.730
Proposed	0.819	0.849	0.799	0.823	25.390	0.889

As shown in Table III, the proposed Modified Haar Wavelet framework consistently outperformed the conventional Haar Wavelet across all evaluation metrics. The IoU increased substantially from 0.350 to 0.819, indicating a significantly higher overlap between the detected White Spot Disease (WSD) markers and the corresponding ground-truth annotations. Similarly, Precision improved from 0.480 to 0.849, demonstrating the effectiveness of the proposed framework in reducing false-positive detections.

The Recall value also increased from 0.550 to 0.799, indicating that a larger proportion of disease markers were successfully detected. As a result, the F1-score improved considerably from 0.510 to 0.823, confirming a better balance between Precision and Recall. These improvements suggest that integrating adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding improves disease marker localization performance compared to the conventional Haar Wavelet approach.

Furthermore, the proposed framework achieved a higher PSNR of 25.390 dB, compared to 19.2 dB with the conventional Haar Wavelet. This result indicates improved preservation of relevant image information while reducing unwanted distortions during detection. Likewise, the SSIM value increased from 0.730 to 0.889, demonstrating superior structural similarity and better preservation of disease-related image characteristics.

Overall, the comparative analysis confirms that the proposed modifications significantly improve the performance of the conventional Haar Wavelet framework. The results demonstrate that integrating adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding improves disease marker detection accuracy, structural preservation, and overall detection reliability.

While the comparative analysis demonstrates that the proposed Modified Haar Wavelet framework significantly outperforms the conventional Haar Wavelet approach, it does not reveal the individual contribution of each component within the proposed framework. Therefore, an ablation study was conducted to systematically evaluate the impact of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding on overall detection performance. By removing one component at a time while preserving the remaining processing stages, the contribution of each component can be quantitatively assessed. The results of the ablation study are presented in Table IV.

Table IV presents the ablation study results of the proposed Modified Haar Wavelet framework. The objective of this analysis is to evaluate the contributions of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding to the overall performance of White Spot Disease marker detection.

TABLE IV. ABLATION STUDY OF THE PROPOSED MODIFIED HAAR WAVELET FRAMEWORK

Configuration	IoU	Prec.	Rec.	F1 Score	PSNR	SSIM
Without Adaptive Coefficient Selection	0.614	0.827	0.599	0.695	19.040	0.844
Without Roberts Operator	0.671	0.833	0.655	0.733	20.818	0.856
Without Hysteresis Thresholding	0.573	0.823	0.559	0.665	17.770	0.835
Proposed	0.819	0.849	0.799	0.823	25.390	0.889

The complete framework achieved the highest performance across all evaluation metrics, obtaining an IoU of 0.819, a precision of 0.849, a recall of 0.799, an F1-score of 0.823, a PSNR of 25.390 dB, and an SSIM of 0.889. These results indicate that integrating all proposed components effectively improves disease marker localization while preserving structural information.

When adaptive coefficient selection was removed, the IoU decreased from 0.819 to 0.614, and the F1-score decreased from 0.823 to 0.695. This reduction demonstrates that adaptive coefficient selection plays an important role in preserving significant wavelet coefficients associated with disease markers while suppressing irrelevant image structures and noise.

Similarly, removing the Roberts operator reduced the IoU to 0.671 and the F1-score to 0.733. The decrease in performance indicates that Roberts operator-based edge enhancement contributes to improved gradient localization and more accurate representation of disease marker boundaries.

The largest performance degradation occurred when hysteresis thresholding was removed. In this configuration, the IoU dropped to 0.573, the F1-score decreased to 0.665, and PSNR and SSIM were reduced to 17.770 dB and 0.835, respectively. These results suggest that hysteresis thresholding plays a critical role in maintaining edge continuity, reducing fragmented detections, and preserving the connectivity of disease marker structures. By linking weak edge responses to neighboring strong edges, hysteresis thresholding enables more complete and coherent localization of disease markers, thereby significantly improving the overall detection performance of the proposed framework.

Overall, the ablation study confirms that each component contributes positively to the final performance of the proposed framework. The superior results achieved by the complete framework demonstrate that the performance improvement is not due to a single processing stage but to the synergistic integration of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding.

To provide a holistic view of detection performance, Fig. 4 presents a visual comparison of all evaluated methods across IoU, Precision, Recall, F1-score, PSNR, and SSIM. This visualization complements the quantitative results by illustrating each method's performance trends and highlighting the consistent superiority of the proposed Modified Haar Wavelet framework across multiple evaluation criteria.

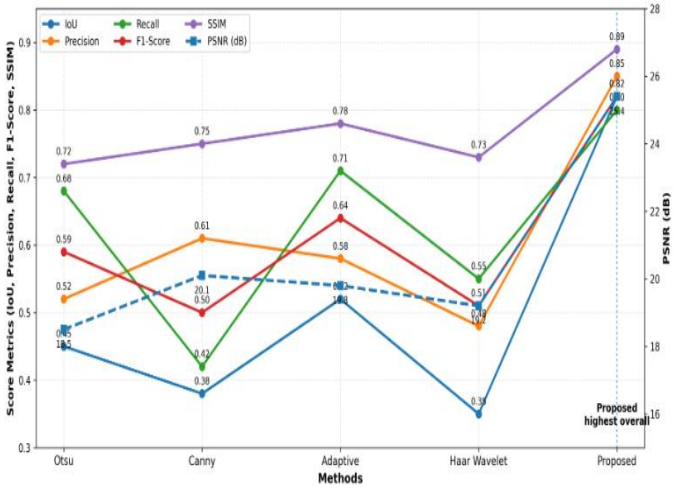


Fig. 4. Overall performance comparison of Otsu, Canny, Adaptive Thresholding, conventional Haar Wavelet, and the proposed Modified Haar Wavelet Framework, evaluated using IoU, Precision, Recall, F1-score, PSNR, and SSIM.

Fig. 4 presents a comprehensive comparison of the evaluated methods across six performance metrics, namely IoU, Precision, Recall, F1-score, PSNR, and SSIM. The graph clearly illustrates the performance trends of each method and highlights the effectiveness of the proposed Modified Haar Wavelet framework.

The proposed method consistently achieved the highest values across all evaluation metrics. In terms of localization accuracy, the proposed framework obtained the highest IoU value of 0.819, indicating superior overlap between the detected White Spot Disease markers and the corresponding ground-truth annotations. Similarly, the highest Precision value of 0.849 demonstrates the framework's ability to minimize false-positive detections while accurately identifying disease markers.

The Recall value of 0.799 indicates that the proposed framework successfully detected most disease markers present in the image. The combination of high Precision and Recall yielded the highest F1-score of 0.823, confirming balanced and reliable detection performance. In contrast, conventional methods such as Otsu, Canny, Adaptive Thresholding, and Haar Wavelet exhibited lower and less consistent performance across these metrics.

The graph also demonstrates the proposed framework's superiority in preserving image quality. The PSNR value reached 25.390 dB, significantly exceeding the values obtained by the conventional methods, which ranged from 18.5 dB to 20.1 dB. Likewise, the proposed framework achieved the highest SSIM value of 0.889, indicating better preservation of structural information and disease-related image characteristics.

Fig. 4 visually confirms the quantitative findings presented in Table I. The consistent dominance of the proposed framework across all evaluation metrics demonstrates that integrating adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding effectively improves

disease marker localization, edge representation, and structural preservation in White Spot Disease detection.

The experimental results further indicate that the proposed Modified Haar Wavelet framework consistently outperforms conventional edge detection methods across all evaluation metrics. The combination of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding significantly improves disease marker localization accuracy, structural preservation, and overall detection reliability for White Spot Disease images.

These findings suggest that the proposed framework effectively prioritizes disease marker localization while preserving important structural information. Although the method achieves superior detection performance, further investigation is required to evaluate computational efficiency, scalability, and real-time applicability in large-scale aquaculture monitoring systems.

V. Conclusion

This study proposed a Modified Haar Wavelet framework for White Spot Disease marker detection by integrating adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding. The proposed framework was designed to improve disease marker localization and edge representation while preserving important structural information in fish images.

Experimental results demonstrated that the proposed framework consistently outperformed Otsu, Canny, Adaptive Thresholding, and conventional Haar Wavelet methods across all evaluation metrics. The proposed method achieved the highest IoU (0.819), Precision (0.849), Recall (0.799), F1-score (0.823), PSNR (25.390 dB), and SSIM (0.889), indicating superior detection accuracy and structural preservation. Comparative analysis further confirmed that the proposed modifications significantly improved the performance of the conventional Haar Wavelet framework.

The ablation study verified the contribution of adaptive coefficient selection, Roberts operator-based edge enhancement, and hysteresis thresholding, demonstrating that the performance improvement resulted from the synergistic integration of all components rather than a single processing stage. These findings highlight the effectiveness of the proposed framework as a robust wavelet-based solution for automated detection of White Spot Disease markers.

Future work will focus on evaluating robustness under more diverse aquaculture conditions, analyzing computational efficiency, and integrating learning-based approaches to further improve detection performance and real-time applicability.

Acknowledgment

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS) [Ref: FRGS/1/2020/ICT02/UUM/02/2, S/O Code: 14856]. The authors alone are accountable for the content presented in this article.

References

[1] M. J. Mia, R. Bin Mahmud, M. S. Sadad, H. Al Asad, and R. Hossain, “An in-depth automated approach for fish disease recognition,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7174–7183, Oct. 2022, doi: 10.1016/j.jksuci.2022.02.023.

[2] A. Waleed, H. Medhat, M. Esmail, K. Osama, R. Samy, and T. M. Ghanim, “Automatic Recognition of Fish Diseases in Fish Farms,” in 2019 14th International Conference on Computer Engineering and Systems (ICCES), 2019, pp. 201–206. doi: 10.1109/ICCES48960.2019.9068141.

[3] S. N. Pauzi, M. G. Hassan, N. Yusoff, N. H. Harun, A. H. Abu Bakar, and B. C. Kua, “A review on image processing for fish disease detection,” in Journal of Physics: Conference Series, IOP Publishing Ltd, Aug. 2021. doi: 10.1088/1742-6596/1997/1/012042.

[4] M. S. Ahmed and S. M. Jeba, “SalmonScan: A novel image dataset for machine learning and deep learning analysis in fish disease detection in aquaculture,” Data Brief, vol. 54, Jun. 2024, doi: 10.1016/j.dib.2024.110388.

[5] A. Rajbongshi, R. Shakil, B. Akter, M. A. Lata, and M. M. A. Joarder, “A comprehensive analysis of feature ranking-based fish disease recognition,” Array, vol. 21, no. December 2023, p. 100329, 2024, doi: 10.1016/j.array.2023.100329.

[6] M. K. Alsmadi and I. Almarashdeh, “A survey on fish classification techniques,” May 01, 2022, King Saud bin Abdulaziz University. doi: 10.1016/j.jksuci.2020.07.005.

[7] O. Ulucan, D. Karakaya, and M. Turkan, “A Large-Scale Dataset for Fish Segmentation and Classification,” in Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ASYU50717.2020.9259867.

[8] K. Buchmann, “Immune response to Ichthyophthirius multifiliis and role of IgT,” Aug. 01, 2020, Blackwell Publishing Ltd. doi: 10.1111/pim.12675.

[9] A. S. B. M. Azhar, N. H. B. Harun, N. B. Yusoff, M. G. Bin Hassan, and K. B. Chu, “Image Enhancement on Underwater Images for Protozoan White Spot Fish Disease Detection,” in 2022 International Conference on Intelligent Systems and Computer Vision, ISCV 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ISCV54655.2022.9806095.

[10] S. Kumar, A. K. Upadhyay, P. Dubey, and S. Varshney, “Comparative analysis for edge detection techniques,” in Proceedings - IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2021, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 675–681. doi: 10.1109/ICCCIS51004.2021.9397225.

[11] P. Srujana, J. Priyanka, V. Y. S. S. S. Patnaikuni, and N. Vejendla, “Edge Detection with different Parameters in Digital Image Processing using GUI,” in Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 795–802. doi: 10.1109/ICCMC51019.2021.9418327.

[12] R. Sun et al., “Survey of Image Edge Detection,” 2022, Frontiers Media SA. doi: 10.3389/frsip.2022.826967.

[13] Y. Wei and M. Xu, “Detection of lane line based on robert operator,” Journal of Measurements in Engineering, vol. 9, no. 3, pp. 156–166, Sep. 2021, doi: 10.21595/JME.2021.22023.

[14] B. Cui and H. Jiang, “An Image Edge Detection Method Based on Haar Wavelet Transform,” in Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 250–254. doi: 10.1109/ICAICE51518.2020.00054.

[15] M. A. Younus and T. M. Hasan, Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform. IEEE, 2020.

[16] Y. Yu, J. Gao, X. Mu, and S. Wang, “Adaptive LSB quantum image watermarking algorithm based on Haar Wavelet transforms,” Quantum Inf. Process., vol. 22, no. 5, May 2023, doi: 10.1007/s11128-023-03926-1.

[17] J. Ramya, H. C. Vijaylakshmi, and H. Mirza Saifuddin, "Segmentation of skin lesion images using discrete wavelet transform," Biomed. Signal Process. Control, vol. 69, Aug. 2021, doi: 10.1016/j.bspc.2021.102839.

[18] A. Kumar, "Study and analysis of different segmentation methods for brain tumor MRI application," Multimed. Tools Appl., vol. 82, no. 5, pp. 7117–7139, Feb. 2023, doi: 10.1007/s11042-022-13636-y.

[19] S. Dey, R. Roychoudhury, S. Malakar, and R. Sarkar, "Screening of breast cancer from thermogram images by edge detection aided deep transfer learning model," Multimed. Tools Appl., vol. 81, no. 7, pp. 9331–9349, Mar. 2022, doi: 10.1007/s11042-021-11477-9.

[20] Q. Sun and Q. Liu, "The Target Fish's Population Detection Based on The Improved Watershed Algorithm," in 2022 7th International Conference on Intelligent Computing and Signal Processing, ICSP 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 507–510. doi: 10.1109/ICSP54964.2022.9778537.

[21] S. Yunhong, Y. Shilei, Z. Xiaojing, and Y. Jinhua, "Edge Detection Algorithm of MRI Medical Image Based on Artificial Neural Network," in Procedia Computer Science, Elsevier B.V., 2022, pp. 136–144. doi: 10.1016/j.procs.2022.10.021.

[22] X. Liu, Z. Yang, and Y. Chai, "A landscape pattern gradient analysis method coupling wavelet algorithm for balancing development and protection activities of the coastal zone: A case study," Ecol. Indic., vol. 165, Aug. 2024, doi: 10.1016/j.ecolind.2024.112218.

[23] X. Zhu and M. Wen, "Image fusion research based on the Haar-like multi-scale analysis," EURASIP J. Adv. Signal Process., vol. 2024, no. 1, Dec. 2024, doi: 10.1186/s13634-024-01118-2.

[24] Y. Li, Z. Liu, J. Yang, and H. Zhang, "Wavelet Transform Feature Enhancement for Semantic Segmentation of Remote Sensing Images," Remote Sens. (Basel)., vol. 15, no. 24, Dec. 2023, doi: 10.3390/rs15245644.

[25] N. You, L. Han, D. Zhu, and W. Song, "Research on Image Denoising in Edge Detection Based on Wavelet Transform," Applied Sciences (Switzerland), vol. 13, no. 3, Feb. 2023, doi: 10.3390/app13031837.

[26] G. Wang, W. Zhao, P. Zou, J. Wang, H. Yin, and Y. Yu, "Quantum image edge detection based on Haar Wavelet transform," Quantum Inf. Process., vol. 23, no. 8, Aug. 2024, doi: 10.1007/s11128-024-04513-8.

[27] W. Huan et al., "Haar Wavelet-Based Classification Method for Visual Information Processing Systems," Applied Sciences (Switzerland), vol. 13, no. 9, May 2023, doi: 10.3390/app13095515.

[28] M. M. Da Rosa, H. B. Seidel, G. Paim, E. A. C. Da Costa, S. Almeida, and S. Bampi, "An Energy-Efficient Haar Wavelet Transform Architecture for Respiratory Signal Processing," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 2, pp. 597–601, Feb. 2021, doi: 10.1109/TCSII.2020.3046919.

[29] H. Kanagaraj and V. Muneeswaran, "Image compression using HAAR discrete wavelet transform," in ICDCS 2020 - 2020 5th International Conference on Devices, Circuits and Systems, Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 271–274. doi: 10.1109/ICDCS48716.2020.243596.

[30] N. S. Dagar and P. K. Dahiya, "Edge Detection Technique using Binary Particle Swarm Optimization," in Procedia Computer Science, Elsevier B.V., 2020, pp. 1421–1436. doi: 10.1016/j.procs.2020.03.353.

[31] F. Orujov, R. Maskeliūnas, R. Damaševičius, and W. Wei, "Fuzzy based image edge detection algorithm for blood vessel detection in retinal images," Applied Soft Computing Journal, vol. 94, Sep. 2020, doi: 10.1016/j.asoc.2020.106452.

[32] R. Bagaria, S. Wadhwani, and A. K. Wadhwani, "A Wavelet Transform and Neural Network Based Segmentation & Classification System For Bone Fracture Detection," Optik (Stuttg)., vol. 236, Jun. 2021, doi: 10.1016/j.ijleo.2021.166687.

[33] P. Rajeswari, K. K. Thyagharajan, V. S. Prabhu, and G. Shree Devi, "Denoising X-Ray Image Using Discrete Wavelet Transform and Thresholding," in Lecture Notes in Electrical Engineering, Springer Science and Business Media Deutschland GmbH, 2022, pp. 191–198. doi: 10.1007/978-981-16-4625-6_19.

[34] C. Guobin, Z. Sun, and L. Zhang, "Road Identification Algorithm for Remote Sensing Images Based on Wavelet Transform and Recursive Operator," IEEE Access, vol. 8, pp. 141824–141837, 2020, doi: 10.1109/ACCESS.2020.3012997.

[35] A. H. Abdel-Gawad, L. A. Said, and A. G. Radwan, "Optimized Edge Detection Technique for Brain Tumor Detection in MR Images," IEEE Access, vol. 8, pp. 136243–136259, 2020, doi: 10.1109/ACCESS.2020.3009898.

[36] J. Gao, B. Wang, Z. Wang, Y. Wang, and F. Kong, "A wavelet transform-based image segmentation method," Optik (Stuttg)., vol. 208, Apr. 2020, doi: 10.1016/j.ijleo.2019.164123.

[37] N. H. Harun, S. N. Md Pauzi, M. G. Hassan, N. Yusoff, B. C. Kua, and N. Gordon, "Integration of contrast-adaptive colour correction and convolutional neural network for cryptocaryon fish disease detection," Aquac. Fish., Jul. 2026, doi: 10.1016/j.aaf.2025.11.008

Adaptive Neuro-Digital Twin with Cross-Domain Multimodal Representation Learning for Early Alzheimer's Disease Prognosis

Dr V S Krushnasamy[1], Dr. Annapurna Mishra[2], Pratik Gite[3], Ganesh Kumar Anbazhagan[4], Dr. Adapa Gopi[5], M.Misba[6], Dr. A. Arul Anitha[7], Osama R.Shahin[8]

Associate Professor, Department of Electronics and Instrumentation Engineering, Dayananda Sagar College of Engineering, Bangalore, India[1]

Associate Professor, Department of Electronics and Communication Engineering, Silicon University, Odisha, India[2]

Associate Professor, Department of Computer Science and Engineering, Parul University- Parul Institute of Technology, Vadodara, Gujarat, India[3]

Department of Microbiology, Saveetha Medical College and Hospital, Saveetha Institute of Medical and Technical Sciences, Chennai – 602 105, Tamil Nadu, India[4]

Associate Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur Dist-522302., Andhra Pradesh, India[5]

Assistant Professor, senior grade, Department of Artificial Intelligence and Machine Learning, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India[6]

Assistant Professor, Department of Computer Applications, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam, Theni Dt. Tamil Nadu, India[7]

Department of Computer Science, College of Computer and Information Sciences, Jouf University, Saudi Arabia[8a]

Physics and Mathematics Department, Faculty of Engineering, Helwan University, Helwan, Egypt[8b]

Abstract*—Alzheimer's disease (AD) refers to a progressive neurodegenerative disease involving cognitive impairment, brain atrophy, and functional neurological deficits that make early prediction and subsequent disease progression monitoring extremely difficult. Currently available AI methods predominantly focus on employing single modality analysis or static multimodal analysis approaches, which tend to solve AD prognosis as a binary classification problem. Also, much of the current literature on AD does not take into account the importance of progression-aware, patient-wise evaluation, and robust management of multimodality data heterogeneity or missingness. To tackle such issues, this study presents the proposal of an Adaptive Neuro-Digital Twin-based framework for predicting Alzheimer's disease, known as ANDT-AD. The proposed framework incorporates the heterogeneous multimodality information in clinical-cognitive data, structural Magnetic Resonance Imaging (MRI), and Electroencephalography (EEG) signals through modality-wise deep encoders. Specifically, a clinical encoder based on a transformer structure is utilized to capture non-linear cognitive interactions, while a Vision Transformer model and an attention-enhanced temporal EEG encoder model help to extract neuroanatomical information from MRI signals and electrophysiological signals in EEG, respectively. The framework is developed using Python and assessed using public domain clinical, MRIs, and OpenNeuro ds004504 EEG data sets for five-fold cross-validation and patient-level evaluations. Experimentation yielded 98.0% diagnostic accuracy with an AUC of 0.97, which exceeds the performance of current multimodal baselines. Moreover, the framework attained an MAE of 1.12, an RMSE of 1.46, and a progression risk C index of 0.89, proving its strength in predicting cognitive decline and personalizing disease progression models under heterogeneity and missingness of multimodal scenarios.

Keywords*—Alzheimer's disease; neuro-digital twin; multimodal learning; disease prognosis; cross-domain representation learning

I. Introduction

AD is an irreversible, progressive neurodegenerative disorder, which is associated with cognitive decline, memory loss, and functional decline, posing substantial clinical, social, and economic challenges worldwide [1], [2]. Due to its protracted preclinical stage, pathological changes, including synaptic impairment, structural atrophy of the brain, and electrophysiological disorders, tend to be present many years before the appearance of clinical symptoms [3], [4]. Nevertheless, the traditional methods of diagnosis are still predominantly reactive and based on cognitive tests and imaging biomarkers that identify disease when extensive neurodegenerative processes have already been experienced [5], [6]. Early prognosis of AD is also complicated by the heterogeneous and multifactorial nature of this disease because the progression of the disease in people differs significantly. As a result, there is an increasing demand for computational models that can combine heterogeneous neurobiological evidence to facilitate early disease stage detection, as well as model disease progression individually [7], [8]. The current innovations in artificial intelligence have shown potential in AD diagnosis, but most of the current models are unchanging, modality-driven, and incapable of incorporating the dynamic changes in neurodegeneration over time.

The current literature has investigated various AI-based methods of analyzing AD in various data modalities. Existing models [9], [10] have been frequently used on MRI scans to classify multiclass stages of AD with an existing high accuracy, with limited interpretability and generalization. Recent studies [11] have enhanced features like global extraction of features in neuroimaging, but still cannot support pipelines that are not based on imaging. The traditional ML models supported by clinical and cognitive datasets, like Support Vector Machines, Random Forests, and gradient-based ensembles, have been examined as they are unable to learn complex temporal dependencies [12], [13]. Existing EEG studies show sensitivity to functional brain variations, but these studies are not usually coupled to structural and clinical history [14], [15]. Although there have been multimodal fusion techniques based on feature concatenation and late fusion techniques, they generally assume patient-aligned data and do not adapt to heterogeneous data and cross-cohort data. In addition, current models consider AD prediction as a single classification problem, which does not consider the dynamic and progressive nature of neurodegeneration, thus limiting its prognostic value.

To address the limitations of existing studies predicting the occurrence of Alzheimer's disease, the proposed study is an attempt to introduce a Progression-Aware Neuro-Digital Twin modeling framework, which describes the neurological status of a patient as a dynamically changing latent state based on multimodal clinical, MRI, and EEG data. In contrast to traditional diagnostic systems based on the unchanging classification, the given solution is capable of recording longitudinal neurodegenerative dynamics and can diagnose a disease, predict cognitive decline, and estimate progress risk in one single learning system.

A. Research Motivation

The task of early detection of disease onset and development is still considered one of the major clinical challenges due to the progressive nature of neurodegeneration and the variability of the disease presentation among different patients. Machine learning-based techniques have been traditionally concerned with only disease stage identification without addressing the issue of patient-specific changes in the neurological states during the development of the disease. Moreover, heterogeneity of multimodal biomedical information, such as clinical, cognitive measurements, brain structural images, and biosignal recordings, restricts the effectiveness of traditional single-modality predictions. Thus, there is an increasing demand for designing a progression-aware computational paradigm capable of handling heterogeneous multimodal information and providing patient-specific disease evolution monitoring using the Neuro-Digital Twins approach.

B. Research Significance

The modeling of Alzheimer's disease, taking into consideration disease progression, can substantially benefit early diagnosis and continuous monitoring of the disease due to the modeling of temporal changes of neurological status instead of only being able to recognize different stages of the condition. Such a novel approach will enable personalized disease trajectory monitoring and risk-sensitive clinical decision making with the help of combining multimodal biomedical information and Neuro-Digital Twin modeling. Continuous synchronization of patient-specific clinical, imaging, and electrophysiological information will allow us to obtain a correct model of the neurodegenerative processes occurring in an individual. This approach could prove highly useful not only for early intervention but also for future intelligent prognosis systems.

C. Problem Statement

While remarkable advancements have been made recently in the field of artificial intelligence in diagnosing Alzheimer's disease [16], most of the currently developed algorithms still employ traditional single modality learning or static multimodal fusion methodologies that presuppose the completeness and synchronicity of the data. These algorithms cannot account for the nonlinear nature and subject-specificity of the brain degeneration process [17]. In clinical practice, different modalities are often missing, corrupted, or asynchronous, which makes it difficult to deploy existing solutions. The failure to accurately combine heterogeneous clinical, imaging, and electrophysiological data while maintaining robustness under missing-modality further limits practical implementation. This study addresses these shortcomings with a progression-conscious adaptive neuro-digital twin architecture for early prognosis of Alzheimer's disease.

D. Key Contributions

- An Adaptive Neuro-Digital Twin (ANDT-AD) paradigm is proposed for early prognosis of Alzheimer's disease based on continuous modeling of the individual neurodegeneration progression pattern.
- A multimodal representation learning method is proposed to merge various types of clinical data, structural MRI images, and EEG signals into a common disease-oriented latent space.
- Specialized modality-specific encoders such as Transformer-based clinical information learning, Vision Transformer-based MRI feature extraction, and enhanced attention-based EEG representation learning are developed.
- Latent State Evolution with progression awareness is formulated to model disease progression patterns and estimate progression risk.
- Experimental studies using publicly available clinical information, MRI, and EEG databases show that the proposed ANDT-AD achieves significantly better performance than other baseline methods.
- The rest of the study is organized as follows: Section II reviews related works on AI-based AD analysis and highlights existing research gaps; Section III explains the proposed adaptive neuro-digital twin methodology; Section IV presents experimental results; and Section V concludes the study with limitations and future research directions.

II. Related Work

Recent works have investigated the use of advanced machine learning and deep learning methods to enhance the diagnosis and

prognosis of Alzheimer's disease (AD). Gryshchuk et al. [18] examined the efficiency of self-supervised learning in the differentiation of neurodegenerative diseases based on T1-weighted MRI data, where no large-scale labeled data are available. Their contrastive learning-based feature extractor with a linear classification head obtained comparable accuracy in balance to the supervised models. Nonetheless, it was a method that depended on a single imaging mode and binary classification, which restricted its application in multimodal integration and fine disease staging.

Similarly, a multimodal representation learning framework proposed by Dai et al. [19] relies on cross-domain distillation to conduct AD classification. The approach involved the use of a pre-trained text encoder to instruct the learning of structural MRI and PET feature extractors, which included MMSE scores by adding positional embeddings. The model performed highly on the ADNI dataset, though it was only evaluated on one benchmark dataset, and it was not adequate in investigating the longitudinal development of the disease.

There are also multimodal fusion methods or techniques that have been studied to enhance the accuracy of diagnosis. Muksimova et al. [20] proposed FusionNet that combines MRI, PET, and CT data through modality-specific encoders and attention-based fusion layers. The framework had a classification accuracy of about 94% that showed the capability to detect AD at an early stage. However, the data augmentation method based on GAN provides an opportunity to overfit, and it is not externally multi-centered.

Cognitive and electrophysiological data have been studied by other researchers. Saleh et al. [16] developed a hybrid CNN–LSTM model using longitudinal established a hybrid CNN-LSTM model based on longitudinal cognitive scores using the method of Bayesian optimization and explainability, with the use of the SHAP and LIME. Although the model enhanced predictive ability, it omitted modalities of imaging, and in that it used a large amount of computational resources. Similarly, Vo et al. [21] suggested a time-frequency representation of an EEG-based deep learning model to screen non-invasive AD. Although the method obtained more than 80% accuracy, it was sensitive to EEG noise, and its generalization to a wide range of neurodegenerative conditions was low.

There has also arisen digital twin and generative modeling. Amato et al. [22] presented a digital twin-based diagnostic model with EEG-derived biomarkers, which showed encouraging prognostic capabilities, though it needs a large-scale validation. Meanwhile, Sharma et al. [23] suggested a generative model to simulate disease progression using both latent diffusion models and neural ODE, but the algorithm requires a lot of computational power.

Finally, Ali et al. [17] focused on the issue of multimodal learning development and suggested a self-supervised model that enhances cross-cohort generalization and uses missing modalities. The method, however, demands a lot of multimodal pretraining data, and the interpretation is limited. In general, existing literature has shown the promise of deep learning, multimodal fusion, and self-supervised learning in diagnosing and prognosing the disease in the case of Alzheimer's. There are, however, several shortcomings that can be seen in the existing approaches.

III. Proposed ANDT-AD Framework for Early Prognosis of Neurodegenerative Diseases

The suggested framework proposes a novel Progression-Aware Multimodal Neuro-Digital Twin framework that utilizes clinical-cognitive evaluations, structural MRI images, and electroencephalographic biosignals in predicting the personalized progression of Alzheimer's disease. In the first step, each modality is independently preprocessed and encoded, thereby producing structural, functional, and behavioral representations relevant to neurodegeneration. The resulting feature embeddings are subsequently aligned through cross-domain multimodal representation learning to form a robust feature space for handling heterogeneous and missing data. In contrast to the existing methods, where modality fusion or progression analysis is independently carried out, the suggested framework concurrently co-trains multimodal representation learning and progression modeling within one framework. The main innovation in this context involves Neuro-Digital Twin, which enables continuous updating of a personalized neurological twin based on multimodal longitudinal measurements, leading to personalized state transition modeling. The workflow of the proposed framework is illustrated in Fig 1.

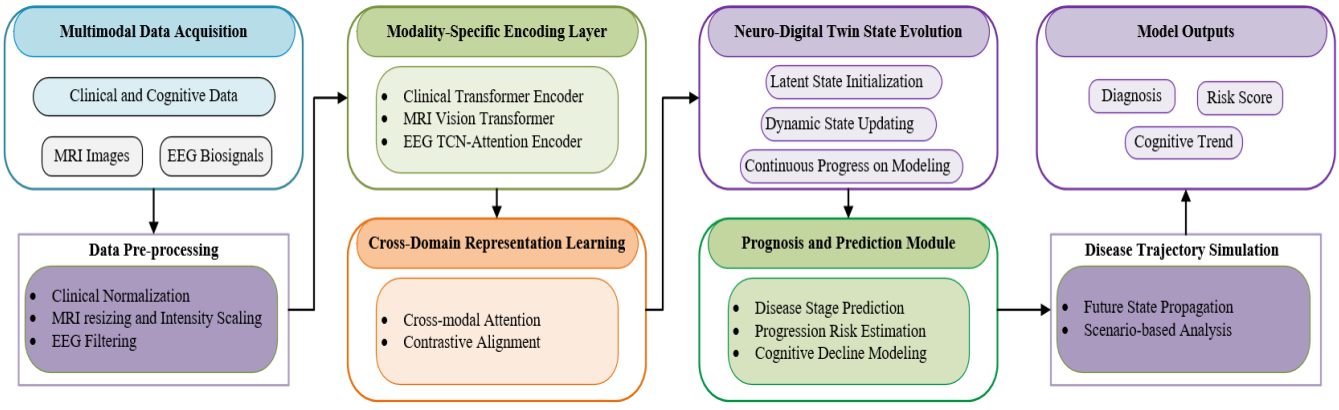


Fig. 1. Overall workflow of proposed ANDT-AD framework.

A. Data Collection and Dataset Description

This study uses three publicly available and supplementary datasets to build a multimodal representation of the AD, which includes clinical, structural, and functional neurological features. The Alzheimer's Disease Dataset [24] It is a clinical and cognitive tabular dataset that is used to offer demographic data, cognitive measuring scores, and clinically significant risk factors that relate to disease development and advancement. The data is an indication of the patterns of high-level cognitive decline and an approximation of regular clinical assessments. The Alzheimer Multiclass Images Dataset [25] that consists of labeled magnetic resonance imaging (MRI) scans of various stages of AD is used to model structural brain changes. With the help of these images, it is possible to extract spatial and morphological biomarkers that are associated with neurodegeneration. Moreover, functional neural dynamics are also measured with the help of the Open Neuro dataset called ds004504 [26] that contains electroencephalography (EEG) measurements that reflect brain activity in terms of time and frequency. The clinical data, MRI data, and EEG data are collected from public datasets and lack full subject correspondence. Consequently, the suggested framework fails to perform patient-wise multimodal fusion. Instead, modality-specific encoders learn separate disease representations that are aligned within a common latent space via cross-domain representation learning. The Neuro-Digital Twin corresponds to a personalized latent neurological state, which undergoes progression-aware state evolution for disease trajectory prediction.

The public Kaggle datasets that were used in this work consist of community-annotated datasets, and the original procedure for clinical annotations of the datasets comes from the dataset providers. The ground-truth annotations were retrieved from the given data set annotations and used as they were. As detailed clinical provenance and annotation protocols are not entirely available for these data sets, this drawback is taken into account when discussing the work. Preprocessing, labeling, quality control, and standard evaluation protocol were used to increase the accuracy of the labels before training the models.

B. Data Preprocessing and Feature Preparation

In order to guarantee consistency across heterogeneous biomedical sources, modality-specific preprocessing pipelines are used to preprocess data from various domains, such as clinical, imaging, and electrophysiological domains, before representation learning. The pre-processing process includes normalization and encoding clinical data, removing noise in the MRI scans, and filtering the EEG data in order to remove any artifacts.

1) Clinical and cognitive data preprocessing: All non-nominal characteristics, such as age, cognitive assessment scores, and indicators that use biomarkers, are standardized using a z-score, which is presented as Eq. (1).

$$x' = \frac{x-\mu}{\sigma} \tag{1}$$

where, μ and σ are the mean and standard deviation of features, respectively. This normalization provides similar scales of features to ensure that high-magnitude variables do not prevail in the process of model training.

2) MRI image preprocessing: The resizing of each scan is done to a constant spatial resolution of 224x224 pixels to ensure that the samples are uniform to one another and that the architecture of the convolutional encoders does not suffer. The min-max scaling is used to normalize the intensity of voxels, averaging to the [0,1] range, and minimize scanner effects on voxel intensity.

3) EEG signal preprocessing: EEG data acquired from the OpenNeuro dataset called "ds004504" are preprocessed to extract clean and useful electrophysiological signal representations. The raw EEG data is filtered using a fourth-order band-pass filter between 0.5 Hz and 45 Hz and a fifth-order notch filter at 50 Hz to eliminate noise, both low- and high-frequencies, as well as power line interference. The ICA technique is used to remove eye movement artifacts and other muscular artifacts in order to preserve brain-related activation patterns. The resulting filtered signals are divided into fixed-length 2-second epochs with 50 percent overlap in order to reflect short-term time changes. The Short-Time Fourier Transform (STFT) method is used to obtain spectral-temporal feature maps for deep learning tasks. Patient-wise stratified sampling is performed in order to avoid a potential data leakage issue where all instances from the same patient fall into only one split, with a ratio of 70 percent for training, 15 percent for validation, and 15 percent for testing.

4) Rationale for selection of encoders: The Transformer model is chosen for modeling clinical features because of its capability of capturing complicated interactions between heterogeneous clinical features by using the self-attention mechanism. ViT is selected for MRI feature extraction as it can learn the global spatial dependency among brain regions, which is crucial for detecting distributed neurodegeneration. TCN is chosen for EEG feature extraction because of its capability of capturing long-range temporal dependency with reduced computational overhead compared to recurrent neural networks. In comparison with the traditional CNN, LSTM, and machine learning models, the chosen encoders show better capability in modeling multimodal features with computational efficiency.

C. Modality-Specific Encoding Architecture

The ANDT-AD is developed as a multimodal latent state-space learning problem. Let the multimodal data be given as in Eq. (2).

$$\mathcal{D} = \{D_c, D_m, D_e\} \tag{2}$$

where, D_c, D_m, and D_e denote the clinical–cognitive dataset, structural MRI dataset, and EEG bio signal dataset, respectively. Both modalities give heterogeneous observations of complementary characteristics of neurological degeneration. The framework aims at learning a parameterized mapping as in Eq. (3).

$$f_\theta: \mathcal{X} \rightarrow \{y_{stage}, r_{risk}, c_{future}\} \tag{3}$$

where, $\mathcal{X}$ denotes multimodal neurological observations, y_{stage} represents disease stage classification, r_{risk} denotes progression risk estimation, and c_{future}corresponds to predicted future cognitive decline. Disease progression is modelled using a persistent latent neuro-digital twin state $s_t \in \mathbb{R}^d$, where s_t represents the hidden neurological condition of a subject at the update step t, and d denotes the latent dimensionality. The proposed Adaptive Neuro-Digital Twin Framework is given in Fig. 2.

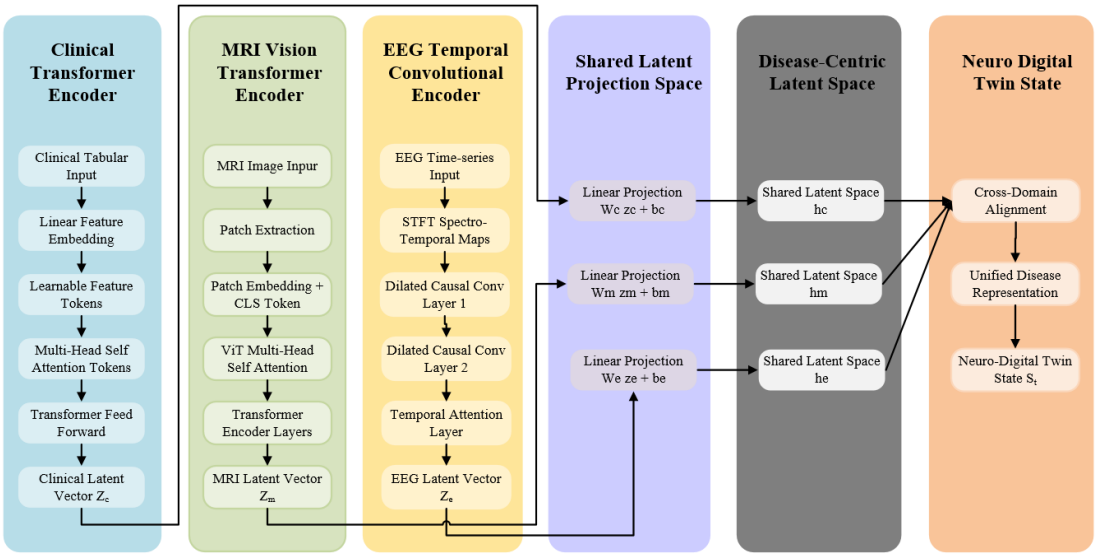


Fig. 2. ANDT-AD framework.

1) Transformer-based encoding of clinical and cognitive data: The clinical and cognitive tabular representations of the patients' data contain vital manifestations of the disease, such as demographic risk factors, neuropsychological evaluation scores, and cognitive impairments. Under the current model architecture, these tabular representations undergo processing using a tabular transformer network encoder that effectively learns the interaction between features in a non-linear manner, an aspect that cannot be achieved by traditional fully connected networks. To begin with, the patient records are encoded as a fixed feature vector representation and projected onto the latent embedding space using a linear transformation. Unlike temporal position encodings, feature embeddings that help retain the identity and meaning of the feature are used. Self-attention techniques are then applied to the embeddings to model context dependencies in clinical attributes and learn their interaction. This can be represented mathematically as shown below:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \quad (4)$$

In Eq. (4), Q, K, and V denote query and key matrices, as well as value matrices, of clinical feature embeddings, and d is an embedding dimension. The encoder produces a very small latent representation of clinically significant patterns related to the progression of Alzheimer's through stacked transformer layers. The performance of this encoder will be an abstracted clinical disease profile, focusing on the severity of risks and cognitive impairment instead of crude clinical values.

Let the multimodal patient input be represented as $X = \{X_c, X_m, X_e\}$, where X_c, X_m, and $X_e$denote clinical records, MRI images, and EEG signals, respectively, providing complementary information for Alzheimer's disease prognosis. The modality-specific embeddings are extracted as $H_c = f_c(X_c)$, $H_m = f_m(X_m)$, and $H_e = f_e(X_e)$, where $f_c(\cdot)$, $f_m(\cdot)$, and $f_e(\cdot)$represent Transformer-based clinical encoding, Vision Transformer-based MRI feature learning, and attention-enhanced temporal EEG modelling. These learned features are fused into a unified latent representation as $Z = \alpha H_c + \beta H_m + \gamma H_e$, where α, β, and γ are adaptive modality weights satisfying $\alpha + \beta + \gamma = 1$. This weighting mechanism enables the framework to prioritize the most informative modality according to patient condition and data quality. The continuously evolving neuro-digital twin state is updated as $S_t = \phi(S_{t-1}, Z_t)$, where $S_t$and S_{t-1}denote the current and previous disease states, $Z_t$is the fused representation at time t, and $\phi(\cdot)$is the nonlinear transition function. Finally, the total optimization objective is defined as $L = L_{cls} + \lambda_1 L_{prog} + \lambda_2 L_{align}$, where L_{cls} denotes diagnostic classification loss, L_{prog} represents disease progression prediction loss, L_{align} ensures cross-modal embedding consistency, and λ_1, λ_2 are balancing coefficients controlling each objective during training.

2) ViT encoding of structural MRI data: MRI images offer anatomical details that can be associated with the progression of Alzheimer's disease, such as cortical atrophy, hippocampus atrophy, and enlarged ventricles. Under this framework, MRI images are subjected to a Vision Transformer (ViT) network architecture in order to learn the global neuroanatomical representation. The individual preprocessed MRI image is subdivided into patches without overlapping, which are flattened and mapped into a high-dimensional space by a

learned linear projection. A classification token is included in the patch sequence to learn the global structure of the disease.

Unlike conventional convolutional networks that primarily capture local spatial hierarchies, the Vision Transformer (ViT) models long-range spatial dependencies through patch-wise self-attention mechanisms. This capability is particularly important in Alzheimer's disease analysis, where distributed brain regions collectively contribute to cognitive impairment. The generated patch embeddings are passed through multiple transformer encoder layers, enabling the model to learn relationships between anatomically distant brain structures. The final representation of the learnable classification token serves as a global feature embedding that summarizes disease-relevant neuroanatomical patterns across the entire brain. The output embedding can be in formally written form as Eq. (5).

$$z_{\text{MRI}} = f_{\text{ViT}}(I) \tag{5}$$

where, I denote the input MRI scan and $f_{\text{ViT}}(\cdot)$ represents the ViT mapping. This latent representation describes neurodegenerative trends at the macro-level and can be considered the structural aspect of the digital twin state.

3) Temporal convolutional encoding of EEG biosignals: To capture the temporal complexity of EEG data, the current study uses a TCN with an attention mechanism. The EEG preprocessed signals (as spectro-temporal feature matrices) are input into a series of expanded causal convolutional layers. The dilated convolution operation is defined as Eq. (6).

$$y(t) = \sum_{k=0}^{K-1} w(k)\, x(t - d \cdot k) \tag{6}$$

where, $x(t)$ being the input signal, $w(k)$ being the convolutional kernel, dis being the dilation factor, and K being the size of the kernel. The TCN is capable of capturing the short-term oscillatory and long-term temporal interdependences with cognitive deterioration by successively enhancing the rate of dilation through the layers. A subsequent attention layer is utilized to put weight on temporally informative portions of the EEG signal to enable the encoder to focus on disease-related neural evolving motions and reduce remaining noise. The resulting EEG latent representation imprints functional electrophysiological features signifying the Alzheimer's pathology.

4) Projection into a shared latent representation space: After modality-specific encoding, the clinical transformer encoder, MRI ViT, and EEG TCN outputs are all functionally transformed into a common latent dimensional space via linear transformation layers. This projection provides a dimensional consistency across modalities and cross-domain alignment. Where "," and "are" used to represent latent embeddings of clinical, MRI, and EEG encoders, respectively. Every embedding is converted to be as Eq. (7).$z_c z_m z_e$

$$h_i = W_i z_i + b_i, i \in \{c, m, e\} \tag{7}$$

where, z_i is the modality-specific latent embedding, W_i is the learnable projection matrix, b_i is the bias vector, and h_i is the projected shared latent representation. The projection parameters W_i and b_i are jointly optimized with encoder parameters through end-to-end backpropagation during training. With this joint optimization, the modality-specific embeddings are converted to a common disease-centric latent representation space while preserving discriminative information.

This latent space can be considered as the state representation of the Neuro-Digital Twin, and it allows incorporating different neurological information into a single format for modeling disease progression. This method is particularly suitable for learning modality-robust representations, as it requires that each encoder learns useful information from its input even when some modalities are not available. In other words, this approach provides resilience to missing-modality cases, cross-cohort differences, and heterogeneous clinical data distributions, making it more generalized and robust in practice.

D. Cross-Domain Representation Learning

Following modality-specific encoding, the proposed model uses a cross-modality representation learning technique that transforms heterogeneous representations into a single disease-aware latent space. The main task of this step is to combine the information derived from both clinical and neuroimaging features without depending on any modality in a way that ensures cross-cohort robustness. Unlike conventional methods that consider the explicit correspondence between individuals, the proposed model does alignment at the representation level, such that embeddings from various modalities can be aligned in a shared latent space by leveraging their similarity in terms of Alzheimer's disease.

The modality weights (α, β, γ) are automatically learned by the system in the process of training, via a learnable fusion layer which uses an attention mechanism. These modality weights are computed by using trainable parameters, and then normalization is done using the softmax function in such a way that the weights satisfy the condition $\alpha + \beta + \gamma = 1$. Backpropagation helps in updating the parameters of the fusion layer using the gradients coming from the optimization loss in the total process. The contribution of each representation can be changed based on its importance for prediction tasks.

Latent embeddings produced by the clinical transformer encoder, MRI ViT, and EEG TCN are initially processed as updated independent semantic tokens of various neurological viewpoints. The model is then used to make cross-modal attention between these embeddings and, based on the relevance of the modality to a disease, actively reweight the contribution of these modalities. The cross-modal attention of a given modality embedding h_i is calculated in terms of other modality embeddings as Eq. (8).

$$\tilde{h}_i = \sum_{j \neq i} \alpha_{ij}\, h_j \tag{8}$$

The attention coefficients are calculated α_{ij} by means of scaled dot-product attention. They reflect how similar in terms of a disease the different modes are. Cross-modal representation learning seeks to build modality-invariant representations that will encode the pathology shared between the manifestations of the same condition. For that purpose, alignment is done via cross-modal attention and contrastive learning.

In order to achieve domain invariance and lessen modality-specific bias, contrastive learning is introduced as an auxiliary alignment goal. The pairs of positive and negative pairs are formed by connecting embeddings corresponding to similar stages of the disease across different modalities and different stages, respectively. The contrastive loss is considered to be Eq. (9).

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\text{sim}(h_i,h_j)/\tau)}{\sum_k \exp(\text{sim}(h_i,h_k)/\tau)} \quad (9)$$

where, $\text{sim}(\cdot)$ denotes cosine similarity between embeddings, h_i and h_j represent embeddings from different modalities, h_k denotes negative samples, τ is the temperature parameter controlling similarity scaling, and L_{con} represents contrastive alignment loss. This loss promotes embeddings similar to the same states of AD to cluster near one another regardless of their originating modality or dataset, and separate dissimilar states.

To deal with incomplete data in the real world, a modality dropout technique was applied during training, in which some or all modalities would be masked at random. It ensured that the model was able to learn solid cross-modal relationships and still achieve good predictions despite the absence of MRI, EEG, or clinical information.

1) Progression-aware neuro-digital twin modeling: The proposed framework also makes use of a Progression-Aware Neuro-Digital Twin modeling scheme, which allows us to track the dynamics of the progression process of Alzheimer's disease. As part of this methodology, the neurological condition of the patients will be modeled via latent twin states based on the aligned embeddings of the multimodal data (clinical, MRI, and EEG). This differs from traditional approaches to classification, wherein the different states associated with disease progression were treated categorically; however, the Progression-Aware Neuro-Digital Twin model will evolve, making it capable of tracking gradual changes to the neurological conditions of the patients. It was expressed in Eq. (10).

$$z_{t+1} = z_t + f(z_t, \Delta t; \theta) \quad (10)$$

where, z_t is the current twin latent state, $z_{t+1}$is a future twin state, $f(\cdot)$ is a progression dynamics function, Δt is a time interval and θ is the model parameter.

E. Neuro-Digital Twin State Evolution

The suggested Neuro-Digital Twin can be considered as a dynamic latent representation of the neurological condition of a patient rather than a static model for disease classification. The initial state of the digital twin is defined using multimodal feature representations, is evolved using temporal state transition, and is validated through diagnosis and prognosis. Adaptation to the individual patient is done using continuous updates of the latent state to reflect the progression of disease over time.

The proposed Progression-Aware Neuro-Digital Twin models Alzheimer's disease progression as a continuous latent trajectory rather than a fixed sequence of disease stages. This enables the framework to capture gradual neurological changes and individual progression patterns. The evolving latent state integrates multimodal representations and provides a dynamic basis for progression risk estimation.

1) Latent state initialization: The latent embedding of the clinical and cognitive data is used to initialize the neuro-digital twin state since these aspects are the most consistent and early predictors of the risk of AD. Let represent the projected clinical-cognitive embedding that results from the modality-specific encoder. The state of the digital twin is initialized with the help of this embedding.$h_c \in \mathbb{R}^d$

$$s_0 = \phi(h_c) \quad (11)$$

In Eq. (11), $\phi(\cdot)$ is a linear transformation, which takes the clinical embedding into the twin state space. This setup is necessary to ensure that the digital twin is initialized with a cognitively based model of disease status, i.e., baseline risk factors and early cognitive impairment, before severe structural or functional loss.

2) Dynamic state updating with multimodal fusion: With further multimodal measurements being accessible, such as structural MRI and EEG-based embeddings, the state of a neuro-digital twin gets revised in a recurring process of state changes. The resulting cross-domain representation learning obtained as the aligned and fused multimodal representation at update is added to the current twin state. The update of the state is determined by Eq. (12).$z_t ts_{t-1}$

$$s_t = \sigma(W_s s_{t-1} + W_z z_t + b) \quad (12)$$

where, W_s and W_z can be learnt, b is a bias value, and $\sigma(\cdot)$ is a nonlinear activation function. This formulation allows the twin state to incorporate novel evidence together with historical context on diseases. The update state is designed so that the changing temporary modality-specific observations are not replaced by the cumulative disease representation and thus encourages temporal stability and awareness of progression.

3) Continuous disease progression modeling: The dynamics of the neuro-digital twin state are developed to reproduce the process of a continuous disease progression. Instead of discrete stage classification, the overt state transitions carry the gradual degradation of the nervous system in the case of AD. In order to implement a smooth temporal evolution, a progression regularization term is added, and it is defined as Eq. (13).

$$\mathcal{L}_{\text{prog}} = \| s_t - s_{t-1} \|_2^2 \quad (13)$$

This limitation promotes an incremental state dynamic that is realistic of neurodegenerative dynamics. The latent state is going to follow a trajectory in the disease-centric embedding space, with the movement and the rate of change.

The diagnosis and prognosis prediction module converts the changing state of Neuro-Digital Twin to meaningful clinical prediction by using the joint prediction framework of diagnostic and prognostic problems. The prediction task is carried out based on the latent representation of Neuro-Digital Twin, which means all the predictions will be made with the consideration of

neurological information that reflects progression and development of diseases, rather than the features from individual modalities. By constantly updating the state of the Neuro-Digital Twin with multimodal information, it can achieve early detection and risk prediction.

4) Early disease stage and progression risk prediction: The neuro-digital twin state is applied to evaluate the current level of the AD at every update step and the risk of disease progression. The head of classification compares the latent state to discrete disease stages, such as cognitively normal, mild cognitive impairment, and AD. This mapping is defined as Eq. (14).$s_t t$

$$\hat{y}_t = \text{softmax}(W_y s_t + b_y) \quad (14)$$

where, W_y and b_y are parameters of learning. Simultaneously, a head of risk estimation can determine a continuous progression risk score, which is an indicator of the possibility of disease progression within a future horizon. The risk score is obtained as Eq. (15).

$$\hat{r}_t = \sigma(W_r s_t + b_r) \quad (15)$$

where, $\sigma(\cdot)$ is the sigmoid activation function. Collectively, these products offer a probabilistic estimate of the existing disease condition and the susceptibility to progression.

5) Cognitive decline trend estimation: To model longitudinal cognitive decline, the framework uses a regression head that predicts future trends in the cognitive scores, given the latent twin state. The equation of the anticipated cognitive path can be written as Eq. (16).

$$\hat{c}_{t+\Delta} = W_c s_t + b_c \quad (16)$$

where, Δ represents the time ahead of prediction. This formulation enables the model to predict the patterns of gradual cognitive decline as opposed to point-wise scores, which provide an insight into the rate of deterioration that may be expected. The regression result is the direct representation of the impact of structural and functional non-brain alterations coded in the twin state. Categorical cross-entropy loss is used to optimize the classification head, binary cross-entropy loss is used to predict progression risk, and MSE loss is used to predict cognitive decline to represent continuous regression behavior.

F. Multi-Task Optimization and Simulation-Based Analysis

The process of diagnosis, risk prediction, and cognitive trend prediction is optimized together via the multi-task learning approach. When simulating the disease trajectory based on its progression analysis, the Neuro-Digital Twin's state is forwardly propagated via the state evolution function. As a result, it becomes possible for the proposed approach to model the dynamics of Alzheimer's disease progression regardless of the presence of observations due to the updating of the latent neurological state. The total loss function is given as Eq. (17).

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{risk} + \lambda_3 L_{reg} + \lambda_4 L_{con} + \lambda_5 L_{prog} \quad (17)$$

where, L_{cls} is the disease stage classification loss, L_{risk} is the progression risk estimation loss, L_{reg} is the cognitive decline regression loss, L_{con} is the contrastive alignment loss, L_{prog} is the progression regularization loss, and λ_i is the task balancing coefficients. In order to improve interpretability in a clinical context, the explainable AI methods have been incorporated into the suggested architecture. The SHAP method has been used to measure the significance of clinical features, including the patient's age, MMSE value, and cognitive measures. The Grad-CAM technique has been used to detect the key brain regions from MRI scans. Moreover, temporal attention maps from EEG signals helped detect significant neurological patterns that influenced the prediction of prognosis.

Algorithm 1: Adaptive Neuro-Digital Twin Framework for Alzheimer's Disease Prognosis

Input:
Multimodal data $D = \{X_c, X_m, X_e\}$
Output:
Disease stage prediction Y_s, progression risk Y_r, cognitive decline estimate Y_c
Step 1: Input multimodal datasets (Xc, Xm, Xe)
Step 2: Preprocess each modality (normalization, filtering, feature preparation)
Step 3: Encode modalities:
Zc ← Clinical Encoder(Xc)
Zm ← MRI Encoder(Xm)
Ze ← EEG Encoder(Xe)
Step 4: Learn shared representation:
Z ← Cross-Domain Fusion(Zc, Zm, Ze)
Progression-Aware Neuro-Digital Twin Learning
Step 5: Update neuro-digital twin state:
St ← Temporal Update(St−1, Z)
Step 6: Predict outcomes:
Ys ← Disease Stage Classifier(St)
Yr ← Progression Risk Predictor(St)
Yc ← Cognitive Decline Regressor(St)
Step 7: Compute joint loss and update model parameters
Step 8: Output (Ys, Yr, Yc)

The entire training and inference process of the proposed ANDT-AD framework is outlined in Algorithm 1. The algorithm comprises encoding the heterogeneous neurological modalities with modality-specific encoders, projection into a common latent space, and cross-domain representation alignment. The aligned embeddings are used to learn a transition function, which is applied to the neuro-digital twin state with an iterative strategy. During inferences, the trained transition dynamics allow forward simulation of disease dynamics to perform prognosis.

ANDT-AD model proposes a novel framework of multimodal progression modeling using clinical, MRI, and EEG data under the Neuro-Digital Twin paradigm for continuous Alzheimer's disease prognosis. In contrast to the existing research works that mainly focus on static classification and multimodal fusion techniques, the proposed framework makes use of cross-modal representation learning to learn modality-invariant latent representations and integrate heterogeneous biomedical data into a common disease-aware feature space. Moreover, the proposed framework uses dynamic Neuro-Digital Twin state evolution for the continuous modeling of disease progression rather than the prediction of individual disease stages. The usage of a clinical Transformer encoder for clinical

features, a vision transformer model for MRI analysis, and an attention model for EEG signals within a unified multitask learning framework further distinguishes the proposed framework from existing multimodal and generative Alzheimer's disease prediction frameworks.

The ANDT-AD framework has several methodological strengths compared to current artificial intelligence techniques employed for predicting the presence of Alzheimer's disease. Unlike other multimodal methods that are based on static fusion or alignment of multimodal data features, the proposed framework uses an adaptive Neuro-Digital Twin representation that learns from multimodal data evidence, including clinical, MRI, and EEG features. The cross-modal representation learning mechanism ensures robust and efficient knowledge transfer between different modalities and is able to address the problem of missing modalities present in the majority of the earlier works. Additionally, the model can perform joint regression and classification learning to estimate disease stage, predict risk of progression, and forecast cognitive decline trajectory. The application of temporal state update makes the proposed approach a progression-aware predictive tool able to account for disease dynamics in each patient. Thus, the introduced framework is an example of a new scalable Neuro-Digital Twin approach that could be applied for personalized disease trajectory analysis.

IV. Result and Discussion

This was carried out through cross-cohort multi-modal analysis involving clinical-cognitive measurements, structural brain imaging, and electroencephalography biosignals. The model developed was intended to be efficient enough to run in real-time clinical settings, taking less than one second on average to predict patient outcomes with the help of GPU optimization. The compact structure of this framework allows it to easily integrate into hospital edge computing devices as well as cloud-based healthcare services.

A. Experimental Setup

Experimental settings are presented here, which are employed to rigorously evaluate the proposed framework. Cross-cohort multimodal data, consisting of clinical-cognitive measurements, structural brain imaging through MRI, and EEG biosignals, were used for evaluation purposes. The proposed architecture is developed in Python, while batch normalization, dropout, and early stopping methods were considered in order to mitigate overfitting. In addition, a five-fold cross-validation approach was used for performance evaluation. Both complete and incomplete cases are considered, and the framework is evaluated based on diagnostic, prognostic, and regression criteria compared against state-of-the-art unimodal and multimodal baselines. The simulation and hardware parameters were given in Table I.

1) Missing modality handling and data splitting strategy: To evaluate the robustness of the proposed framework under incomplete data conditions, different missing-modality scenarios were considered, including clinical-only, MRI-only, EEG-only, and partial multimodal combinations. The model performance was evaluated by removing individual modalities while maintaining the shared latent representation learning capability. For data partitioning, patient-wise splitting was applied to ensure that samples from the same subject were not distributed across training, validation, and testing sets. This strategy prevents data leakage and provides a reliable assessment of the model's generalization capability.

TABLE I. Simulation and Hardware Setup

Parameter	Configuration
Programming framework	Python with PyTorch
DL libraries	PyTorch Lightning, NumPy
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	16
GPU	NVIDIA RTX 3080 (10 GB)
CPU	Intel Core i9
RAM	32 GB
Training and inference	GPU-accelerated

TABLE II. Model Architecture and Implementation Parameters

Parameter	Configuration
Latent Embedding Dimension	256
Clinical Encoder	Transformer-based encoder (4 layers, 8 attention heads)
MRI Encoder	Vision Transformer (ViT) with 16 × 16 patch size
EEG Encoder	Temporal Convolutional Network (TCN) with dilation rates (1, 2, 4, 8)
Multimodal Fusion Method	Cross-modal attention-based fusion
Dropout Rate	0.2
Optimization Settings	Adam optimizer with learning rate 1×10^{-4}

The implementation parameters for the suggested framework are listed in Table II. Latent embedding dimension refers to the common feature space utilized by multimodal representation learning. Transformers, Vision Transformers (ViT), and Temporal Convolutional Network (TCN) are employed as feature extractors of clinical, MRI, and EEG modality, respectively. The cross-modal attention network incorporates the representations obtained through transformers, ViT, and TCN. Finally, dropout is applied to mitigate overfitting, and Adam with a learning rate of 1×10^{-4} facilitates efficient training.

B. Multimodal Encoding Performance Analysis

The efficiency of the domain-specific encoders utilized by the proposed framework is assessed before proceeding with cross-domain alignment. The analysis evaluates the ability of each encoder in capturing disease-related information within its corresponding data stream and discriminating between different Alzheimer's disease stages within the learned latent spaces. The chosen encoder network architectures are additionally supported by a comparative performance assessment and visualization of the learned latent space, ensuring solid ground for future multimodal integration and Neuro-Digital Twin development. The findings correspond well with clinical knowledge,

suggesting overlapping characteristics of early Alzheimer's disease within cognitive and neurological domains.

C. Cross-Domain Representation Alignment Results

In this sub-section, an assessment of the efficacy of the designed cross-modal representation learning paradigm for aligning heterogeneous modality-specific embeddings within a unified latent space that represents diseases is presented. It is analyzed in terms of how invariant across domains the aligned latent representations turn out to be, the stability of their acquisition during the contrastive optimization process, and, consequently, their improved discriminative power with respect to the considered diseases. In the case of Alzheimer's disease, it is confirmed that the latent representations are indeed effective.

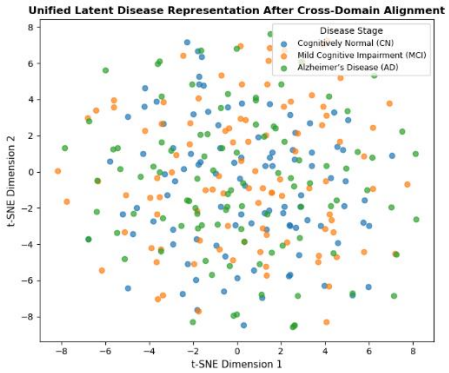


Fig. 3. t-SNE visualization of unified latent space cross-domain alignment.

Fig. 3 shows the t-SNE projection of the unified latent representations obtained following the implementation of the suggested cross-domain alignment mechanism on a two-dimensional projection. The aligned latent space has significantly more pronounced separation between cognitively normal, mild cognitive impairment, and AD categories, and less overlap between the disease stages. This increased clustering is a sign that modality-specific biases were effectively reduced and that the learned representation represents disease-centric features not affected by the data source. This alignment results in a consistent latent disease space that forms a predictable basis of downstream digital twin state evolution and prognostic model, and it will be able to have consistent interpretation across different modalities.

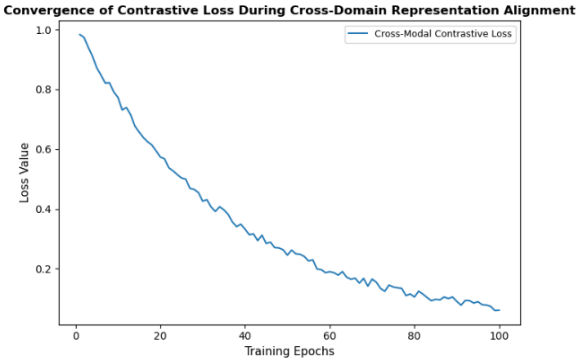


Fig. 4. Contrastive loss convergence and alignment stability.

Fig. 4 shows the convergence pattern of the cross-modal contrastive loss when training the representation alignment module. The decay rate through the epochs is smooth and stable, and this indicates that the optimization and the reduction of inter-modality discrepancies in the latent space are effective. The lack of oscillatory or divergent trends means that the training process is stable and well-built.

D. Progression-Aware Neuro-Digital Twin State Evolution Analysis

The suggested approach proposes a Progression-Aware Neuro-Digital Twin modeling technique where the state of the neurological system is modeled as a constantly evolving latent state. While traditional classification models estimate a discrete state of the disease, the Neuro-Digital Twin model applies temporal state modeling where the neurological state evolves. With this approach, the progression of the disease can be considered for modeling Alzheimer's disease dynamics and generating personalized disease trajectories.

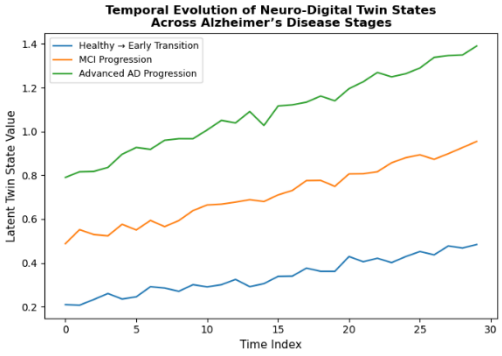


Fig. 5. Temporal evolution of neuro-digital twin states.

Fig. 5 visualizes the dynamic change of the latent states of the Neuro-Digital Twin throughout time in the presence of various states of the AD development. The continuous and slowly rising curves are indicative of the fact that the model is capable of representing smooth neurodegenerative changes as opposed to stage transitions. The unique developmental patterns in healthy, mild cognitive impairment, and Alzheimer's states are evidence that the digital twin stores the severity of diseases in a time-consistent way.

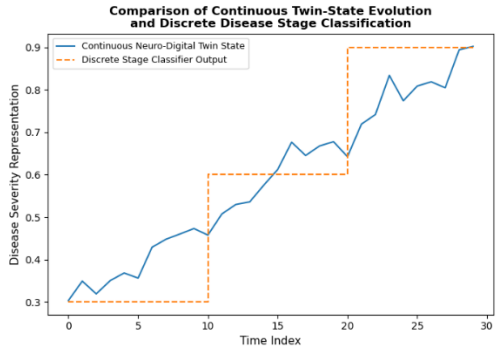


Fig. 6. Continuous twin-state transition vs discrete classification.

Fig. 6 contrasts the Neuro-Digital Twin state development suggested with the results of a traditional discrete-stage classifier. The continuous twin curve has smooth transitions, which are consistent with the smooth process of AD progression, and the discrete classifier creates a sudden step

transition. Such a breakdown is an indication of the weakness of the statistical classification methods in the model of longitudinal disease.

E. Prognosis and Cognitive Decline Prediction Performance

This sub-section provides an analysis of the diagnostic performance of the Neuro-Digital Twin architecture that extends the concept of prognosis beyond the scope of conventional diagnostic classification. The analysis will center on the prediction of trends in cognitive decline longitudinally within the framework of multi-task learning for optimization of diagnostic and prognostic functions. Accuracy and AUC will be used as metrics for diagnostic assessment, whereas prognostic evaluation will consider the prediction of disease trajectory.

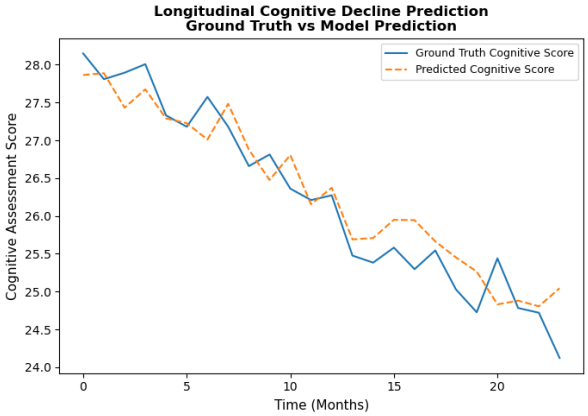


Fig. 7. Longitudinal cognitive score prediction vs ground truth.

Comparison of the ground truth cognitive test scores with the longitudinal prediction scores made by the suggested Neuro-Digital Twin approach can be seen in Fig. 7. Good correlation between the predicted scores and actual cognitive changes indicates the ability of the model to detect gradual change in cognition and not merely detect disease states at different time points. It is evident from the results that there is a good ability of the proposed approach for predicting Alzheimer's disease in the early stages of the disease.

TABLE III. QUANTITATIVE PERFORMANCE OF DIAGNOSTIC AND PROGNOSTIC TASKS

Task	Metric	Proposed ANDT Framework	Best Baseline
Disease stage diagnosis	Accuracy (%)	98.0	92.1
Disease stage diagnosis	AUC	0.97	0.93
Cognitive score prediction	MAE	1.12	1.85
Cognitive score prediction	RMSE	1.46	2.10
Progression risk estimation	C-index	0.89	0.81

The framework performance is quantitatively summarized in Table III by applying it to the diagnostic classification, cognitive decline prediction, and the progression risk estimation tasks. Its persistently high accuracy and AUC values show that it is reliable in disease stage discrimination, whereas lower MAE and RMSE show that it is more accurate in predicting cognitive scores.

F. Disease Trajectory Simulation and Risk Forecasting

This section assesses the capacity of the proposed Neuro-Digital Twin system to predict the future development of Alzheimer's disease. In particular, the proposed framework provides personalized predictions of Alzheimer's disease progression by simulating future developments using the learnt twin latent states through the time-based state evolution process. The model utilizes these future developments to predict future progression of the disease, as well as provide future neurodegeneration predictions.

G. Ablation Study

In this section, the contribution of each major architectural component in the proposed framework is strictly evaluated using a controlled ablation analysis. The study quantifies the effect of each of the four important modules, cross-domain alignment, Neuro-Digital Twin evolution, EEG modality, and contrastive learning, by systematic removal of each one of them in diagnostic performance.

The process of ablation analysis has been additionally applied for the evaluation of the contribution of each of the main components of the framework that was introduced in this study. Alongside the assessment of diagnostic accuracy, the contribution of each module regarding the risk prediction of disease progression is also assessed through the evaluation of changes in the representation of Neuro-Digital Twin states and the ability to estimate the progression of the disease. The absence of cross-domain alignment is used for the evaluation of the importance of multimodal feature synchronization, while the exclusion of Neuro-Digital Twin evolution is evaluated as the impact of temporal disease progression modeling.

TABLE IV. ABLATION STUDY OF KEY ARCHITECTURAL COMPONENTS

Configuration	Accuracy (%)	AUC
Proposed (Full Model)	98.0	0.992
Without Cross-Domain Alignment	93.4	0.951
Without Twin Evolution	92.6	0.944
Without EEG Modality	90.8	0.928
Without Contrastive Learning	91.9	0.936

As described in Table IV, the quantitative assessment analysis on the effects of ablation indicates the accuracy and AUC achieved from various architectural configurations for the proposed approach. Without the use of the cross-domain alignment, there was a 4.6% drop in accuracy, and without using the Neuro-Digital Twin evolution, there was a 5.4% drop in accuracy. This shows the significance of using the temporal aspect of disease progression in addition to multimodal fusion. The lack of contrastive learning in the architecture leads to a drop in cross-modal consistency, hence poor discriminability. Further, omission of the twin-state evolution mechanism leads to less prognosis sensitivity because there is no modeling of disease progression. Removal of the EEG-related aspects of the framework limits multimodal complementarity.

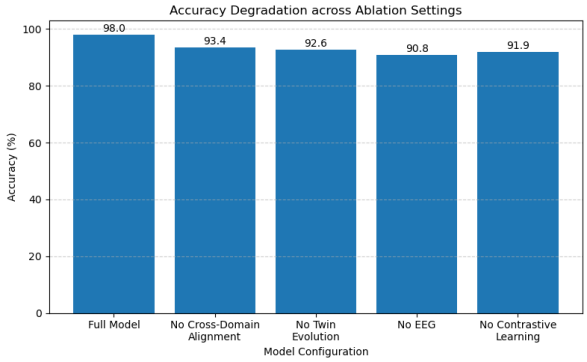


Fig. 8. Accuracy degradation across ablation settings.

Fig. 8 shows how the deletion of the main architectural elements affects the accuracy of the diagnoses, which measures the contribution of each of them to the suggested framework. Maximum decrease is found when the alignment of cross-domain representations is omitted, which emphasizes its importance in the learning of modality-invariant disease patterns. This also reduces accuracy with the exclusion of the Neuro-Digital Twin evolution module, which means that modeling continuous disease dynamics can be used as a complement to the model instead of a substitute for the model.

H. Robustness Under Missing-Modality Conditions

This subsection studies the validity and robustness of the proposed approach when faced with realistic missing modality settings, which can happen in practice owing to various reasons such as financial constraints, availability issues, or non-compliance on the part of the patients. The effect of missing modality on the accuracy of the diagnosis will be tested, and in turn, prove the validity of the approach based on the modality robust representation learning and domain adaptation used in the proposed framework. It can be seen from the results of experiments that the proposed model shows consistent prediction despite varying and incomplete datasets.

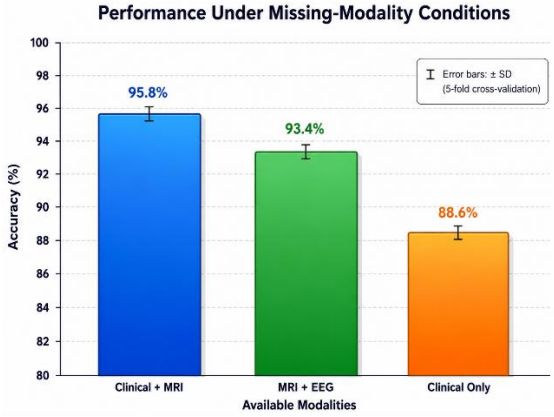


Fig. 9. Model performance under missing-modality scenarios.

Fig. 9 presents an evaluation of the proposed ANDT-AD system for various missing-modality cases. The figure demonstrates how accuracy would be affected by various possible combinations of multi-modal input. For the case when both Clinical and MRI inputs are considered, the highest value of 95.8% is obtained, proving that these two modalities share strong complementation in the sense of cognitive versus structural neuroimaging data. Even for the situation where only MRI and EEG data can be accessed, the system still attains a relatively high value of 93.4%, which proves the efficacy of the cross-modal representation learning and modality-robust fusion approach. The case for only Clinical inputs produces an accuracy rate of 88.6%, reflecting some loss in performance, yet retaining reasonable ability for prediction. The slight loss in performance reflects well on the robustness of the proposed Neuro-Digital Twin system against missing modality. It is worth noting that the system also performs consistently, as reflected by the low standard deviations shown by error bars.

I. Comparative Analysis

A comparative study of the ANDT architecture against existing multimodal Alzheimer's disease prediction models is shown in this subsection. The comparison reveals the results obtained through the use of different types of data as well as their integration. It illustrates the efficiency of each approach through the results gained in terms of diagnostic performance. Thus, it shows how important it is to integrate clinical, imaging, and EEG data by means of cross-domain representation learning and Neuro-Digital Twin modeling in order to predict Alzheimer's disease more accurately and robustly.

TABLE V. PERFORMANCE COMPARISON

Study	Modalities	Accuracy (%)	Precision (%)	Recall (%)
Gryshchuk et al. [18]	T1-weighted MRI	88	87	86
Muksimova et al. [20]	MRI, PET, CT	94	93	92
Vo et al. [21]	EEG (spectrograms, scalograms, Hilbert spectra)	80	79	78
Amato et al. [22]	EEG	88	88	87
Sharma et al. [23]	MRI, cognitive scores, biomarkers	88	87	88
Ali et. Al [17]	MRI, PET	93	92	93
Proposed ANDT-AD	Clinical, MRI, EEG	98	**97**	**97**

Table V shows that the proposed ANDT framework performs better than the existing multimodal Alzheimer models both in terms of diagnostic accuracy and the level of methodological sophistication. Whereas other previous research works using one of the above modalities (MRI or EEG) have an accuracy of between 80 and 94 %, ANDT has an accuracy of 98 % with the use of complementary data, which includes clinical, MRI, and EEG data. Representation alignment and alignment of cross-domain guarantee the existence of modality-invariant latent embeddings, and Neuro-Digital Twin state evolution and multi-task learning together achieve the joint task optimization of diagnostic and prognostic tasks. The findings prove that the implemented framework offers a more trustworthy and clinically relevant forecast over the traditional fusion or generative methods, which demonstrates its effectiveness and practical applicability in the early signs of a disease prognosis.

ANDT-AD approach was compared with a number of competing baselines such as CNN, LSTM, Vision Transformer, Multimodal Transformer, and Graph Neural Network approaches. Experiments revealed the better performance of the proposed approach in terms of classification accuracy and lower AUC errors. These comparative results in Table V must be considered with respect to differences in the dataset used, modality availability, and experimental conditions of the experiments being compared. The proposed framework employs the use of multimodal representations of clinical data, MRI, and EEG data, while some of the approaches already in existence make use of single or dual modalities in independent datasets. These performance gains are not, therefore, due to the superior architecture alone but rather used for comparison purposes to illustrate how the approach works.

J. Statistical Validation

For assessing statistical significance, a five-fold cross-validation was done for all sets. Mean values are used along with their standard deviations for reporting the results. Statistical significance of the better performance obtained with the suggested model was validated with a paired t-test with a significance level of $p<0.05$.

Five-fold cross-validation is performed to determine the reliability and stability of the proposed approach. During every fold, the data is partitioned into five parts that undergo training and testing phases. The performance reliability is measured through average and standard deviation of the accuracy, precision, recall, F1-score, and AUC scores. Moreover, the statistical test performed on the results obtained from a comparison between the proposed approach and the best multimodal baseline approach validates the effectiveness of the proposed method.

K. Analysis of Reliability of the Proposed Framework

In order to assess the robustness and reliability of the proposed framework, several experimental sessions with varying random seed numbers have been carried out. Results of performances are reported with their average values, corresponding standard deviations, and confidence intervals. Significance of the results has been tested statistically by comparing with baselines with a significance value of $p < 0.05$.

TABLE VI. STATISTICAL VALIDATION USING 5-FOLD CROSS-VALIDATION

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	AUC
Fold 1	97.6	96.8	96.2	0.96	0.96
Fold 2	98.1	97.4	97.0	0.97	0.97
Fold 3	97.9	97.1	96.8	0.97	0.97
Fold 4	98.3	97.6	97.2	0.97	0.97
Fold 5	98.0	97.3	96.9	0.97	0.97
Mean	98.0	97.2	96.8	0.97	0.97
Std. Dev.	0.26	0.29	0.34	0.01	0.01

Table VI shows the cross-validation performance results of the proposed multimodal framework in five folds. Each fold has metrics of accuracy, precision, recall, F1-score, and AUC. The average and standard deviation of these values depict the consistency in the performance of the model, which implies the enhancement of the consistency in the generalization of the model and the predictive accuracy in the case of heterogeneous multimodal datasets.

TABLE VII. CROSS-DATASET GENERALIZATION PERFORMANCE

Training Dataset(s)	Testing Dataset	Modality Setup	Accuracy (%)	AUC	MAE	RMSE
Clinical + MRI	EEG	Cross-modality	94.3	0.94	1.45	1.78
MRI	Clinical Dataset	Cross-domain	93.6	0.93	1.52	1.85
EEG	MRI Dataset	Cross-domain	92.8	0.92	1.61	1.92
Clinical + EEG	MRI Dataset	Partial multimodal	95.1	0.95	1.34	1.70
Clinical + MRI + EEG	Independent Hold-out Set	Full multimodal	96.2	0.96	1.28	1.63

Table VII shows the cross-datasets generalization results of the proposed ANDT-AD approach. The framework was trained using one or several source datasets and tested using target datasets with different modalities. Even under varying domains, the framework retains its effectiveness with accuracy levels above 92%. The marginal decline in accuracy is evident when comparing the performance of this framework between in-dataset and cross-dataset settings, which indicates the effectiveness of the proposed approach.

L. Discussion

The experimental results show that the designed Progression-Aware Neuro-Digital Twin framework has performed better than the existing frameworks for diagnosis as well as prognostication of Alzheimer's disease. The framework was able to perform a diagnostic prediction with an accuracy of 98.0% along with an AUC value of 0.97, which is substantially higher than the accuracy level of the strongest multimodal baseline framework at 92.1%. As such, the effectiveness of multimodal representation learning and cross-domain alignment has been proven through the experiment. For longitudinal cognitive decline prediction, the presented approach was able to yield smaller regression errors of MAE=1.12 and RMSE=1.46 compared to baseline error scores of MAE=1.85 and RMSE=2.10, thereby providing stronger predictive performance for modeling temporal neurological progression. Additionally, the obtained C-index for progression-risk estimation showed a score of 0.89, suggesting a high-level ability of individualized prediction of disease progression trajectory and increased sensitivity to progression risk. From the ablation study, it can be concluded that each architectural design element contributed differently to prediction outcomes. Specifically, exclusion of the cross-domain alignment module led to 93.4% accuracy, and exclusion of the Neuro-Digital Twin state evolution module yielded 92.6% accuracy.

V. CONCLUSION AND FUTURE WORK

A novel Adaptive Neuro-Digital Twin (ANDT-AD) framework was proposed to achieve early prediction and progression-informed monitoring of Alzheimer's disease based

on robust multimodal representation learning. Specifically, the proposed framework jointly utilizes information from clinical-cognitive data, structural MRIs, and EEG biosignals by embedding them into a comprehensive latent space associated with disease progression by virtue of cross-domain representation learning and adaptive Neuro-Digital Twin modeling. Contrary to existing static approaches to classification, the proposed method takes into account the individual neurological state evolution process, thus making disease progression prediction possible. Results of experiments demonstrated the superiority of the proposed method compared to the state-of-the-art techniques due to the application of multimodal learning, modeling of neurological states evolution, and Neuro-Digital Twin modeling. In particular, the proposed method proved very successful in longitudinal analysis of disease progression and cognitive deterioration in order to support decision-making in clinical practice.

The future research will address issues related to the development of the suggested methodology for privacy-preserving and large-scale collaborative healthcare environments by using federated learning approaches to train multiple hospitals. Architectures based on graph neural networks might be used to develop models for effective analysis of functional and structural brain connectivity for improved representation learning. In addition, generative models can be developed to generate synthetic data for cases when some modalities are missing. Future research will further involve real-time monitoring of patients using wearable and IoT-based neurological sensors in order to continuously update Neuro-Digital Twins and increase the prediction accuracy. Explainable artificial intelligence methods might also be employed to facilitate the process of clinical interpretation.

REFERENCES

[1] S. K. G.U., A. Purushothaman, P. R., and S. K.G., "AlzFusionFormer: Integrating multiple transformers for early Alzheimer 's disease detection from multi-modal data," Biomedical Signal Processing and Control, vol. 112, p. 108601, Feb. 2026, doi: 10.1016/j.bspc.2025.108601.

[2] I. Kaur and R. Sachdeva, "Prediction Models for Early Detection of Alzheimer: Recent Trends and Future Prospects," Arch Computat Methods Eng, vol. 32, no. 6, pp. 3565–3592, Aug. 2025, doi: 10.1007/s11831-025-10246-3.

[3] S. De Keulenaer et al., "Diagnostic utility of electrophysiological markers for early and differential diagnosis of Alzheimer's, Frontotemporal, and Lewy Body dementias: A systematic review," Neurol Sci, vol. 46, no. 9, pp. 4103–4120, Sep. 2025, doi: 10.1007/s10072-025-08207-6.

[4] A. Sohrabpour, D. Sarovic, H. Zetterberg, M. S. Hämäläinen, S. Baillet, and S. Khan, "Tracking electrophysiological signatures of Alzheimer's disease: a systematic review of multimodal studies," Alzheimer's & Dementia, vol. 21, no. 11, p. e70835, Nov. 2025, doi: 10.1002/alz.70835.

[5] D. R. Thal, K. Poesen, R. Vandenberghe, and S. De Meyer, "Alzheimer's disease neuropathology and its estimation with fluid and imaging biomarkers," Mol Neurodegeneration, vol. 20, no. 1, p. 33, Mar. 2025, doi: 10.1186/s13024-025-00819-y.

[6] R. Jamal et al., "Key biomarkers in Alzheimer's disease: Insights for diagnosis and treatment strategies," Journal of Alzheimer's Disease, vol. 105, no. 3, pp. 679–696, Jun. 2025, doi: 10.1177/13872877251330500.

[7] V. S. Pimprale and H. D. Gadade, "The Role of Artificial Intelligence in the Diagnosis of Alzheimer's Disease - A Comprehensive Review," in 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS), Prawet, Thailand: IEEE, Mar. 2025, pp. 720–725. doi: 10.1109/ICMLAS64557.2025.10968129.

[8] L. Wang, J. Glass, L. Kourtis, and R. Au, "Multi-modal data analysis for early detection of alzheimer's disease and related dementias," The Journal of Prevention of Alzheimer's Disease, vol. 13, no. 1, p. 100399, Jan. 2026, doi: 10.1016/j.tjpad.2025.100399.

[9] D. K. Moorthy, P. Chinnasamy, and P. Nagaraj, "Optimization enabled ResNet features with transfer learning for Alzheimer's disease detection," Computational Biology and Chemistry, vol. 119, p. 108613, Dec. 2025, doi: 10.1016/j.compbiolchem.2025.108613.

[10] S. B. Francis and J. Prakash Verma, "Deep CNN ResNet-18 based model with attention and transfer learning for Alzheimer's disease detection," Front. Neuroinform., vol. 18, p. 1507217, Jan. 2025, doi: 10.3389/fninf.2024.1507217.

[11] Md. A. Mahmud Joy, S. Nasrin, A. Siddiqua, and D. Md. Farid, "ViTAD: Leveraging modified vision transformer for Alzheimer's disease multi-stage classification from brain MRI scans," Brain Research, vol. 1847, p. 149302, Jan. 2025, doi: 10.1016/j.brainres.2024.149302.

[12] M. Kumar, B. Kumar, P. Sharma, R. Sharma, M. Al-Dhaifallah, and A. Shakoor, "Attentive deep learning with Randomized Vector Energy Least Square Twin Support Vector Machine for Alzheimer's Disease diagnosis," Computers and Electrical Engineering, vol. 126, p. 110412, Aug. 2025, doi: 10.1016/j.compeleceng.2025.110412.

[13] M. Moraveji and N. Mansouri, "Recent Advances in Computational and Machine-Learning Approaches for Alzheimer's Disease Classification: A Comprehensive Review," Arch Computat Methods Eng, Dec. 2025, doi: 10.1007/s11831-025-10476-5.

[14] W. Khan et al., "An explainable and efficient deep learning framework for EEG-based diagnosis of Alzheimer's disease and frontotemporal dementia," Front. Med., vol. 12, p. 1590201, Jul. 2025, doi: 10.3389/fmed.2025.1590201.

[15] M. Acharya, R. C. Deo, P. D. Barua, A. Devi, and X. Tao, "EEGConvNeXt: A novel convolutional neural network model for automated detection of Alzheimer's Disease and Frontotemporal Dementia using EEG signals," Computer Methods and Programs in Biomedicine, vol. 262, p. 108652, Apr. 2025, doi: 10.1016/j.cmpb.2025.108652.

[16] H. Saleh, N. ElRashidy, M. Abd Elaziz, A. O. Aseeri, and S. El-Sappagh, "Genetic algorithm-based hybrid deep learning model for explainable Alzheimer's disease prediction using temporal multimodal cognitive data," Int J Data Sci Anal, vol. 20, no. 2, pp. 1073–1103, Aug. 2025, doi: 10.1007/s41060-024-00514-z.

[17] S. B. Ali, N. E. Ghannam, H. Mancy, and B. Gh. Elkilany, "Multimodal Self-Supervised Learning for Early Alzheimer's: Cross-Modal MRI–PET, Longitudinal Signals, and Site Invariance," Diagnostics, vol. 15, no. 24, p. 3135, Dec. 2025, doi: 10.3390/diagnostics15243135.

[18] V. Gryshchuk, D. Singh, S. Teipel, M. Dyrba, and the ADNI, AIBL, FTLDNI study groups, "Contrastive self-supervised learning for neurodegenerative disorder classification," Front. Neuroinform., vol. 19, p. 1527582, Feb. 2025, doi: 10.3389/fninf.2025.1527582.

[19] Y. Dai et al., "Strong Multimodal Representation Learner through Cross-domain Distillation for Alzheimer's Disease Classification," in 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal: IEEE, Dec. 2024, pp. 3144–3149. doi: 10.1109/BIBM62325.2024.10822481.

[20] S. Muksimova, S. Umirzakova, J. Baltayev, and Y. I. Cho, "Multi-Modal Fusion and Longitudinal Analysis for Alzheimer's Disease Classification Using Deep Learning," Diagnostics, vol. 15, no. 6, p. 717, Mar. 2025, doi: 10.3390/diagnostics15060717.

[21] T. Vo, A. K. Ibrahim, and H. Zhuang, "A Multimodal Multi-Stage Deep Learning Model for the Diagnosis of Alzheimer's Disease Using EEG Measurements," Neurology International, vol. 17, no. 6, p. 91, Jun. 2025, doi: 10.3390/neurolint17060091.

[22] L. G. Amato et al., "Digital twins and non-invasive recordings enable early diagnosis of Alzheimer's disease," Alz Res Therapy, vol. 17, no. 1, p. 125, May 2025, doi: 10.1186/s13195-025-01765-z.

[23] R. Sharma, V. Kukreja, S. Hariharan, and S.-Y. Chen, "cLDM-ODE: A Multimodal Generative Framework for Uncertainty-Aware Forecasting of Alzheimer's Disease Progression," IRBM, vol. 47, no. 1, p. 100926, Feb. 2026, doi: 10.1016/j.irbm.2025.100926.

[24] [Rabie El Kharoua, "🧠 Alzheimer's Disease Dataset 🧠." Accessed: Jun. 23, 2026. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset

[25] Aryan Singhal, "Alzheimer's Disease Multiclass Images Dataset." Accessed: Jun. 23, 2026. [Online]. Available: https://www.kaggle.com/datasets/aryansinghal10/alzheimers-multiclass-dataset-equal-and-augmented

[26] Andreas Miltiadous, OpenNeuroDatasets/ds004504. (May 29, 2026). OpenNeuroDatasets. Accessed: Jun. 23, 2026. [Online]. Available: https://github.com/OpenNeuroDatasets/ds004504

Semantic Style Transfer for Paintings Using Convolutional Neural Networks (CNNs)

Hafiz Muhammad Jamsheed Nazir, Zheng Jiangbin, Omar Alsaleh
Department of Software, Northwestern Polytechnical University, Xi'an, China[1, 2]
College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia[3]

***Abstract*—In recent years, the importance of photographic portrait styles has garnered significant attention, prompting numerous researchers to explore innovative methods for modifying and enhancing these styles. Neural style transfer has advanced rapidly for photographic portraits, yet transferring painterly styles to human facial and body images remains difficult because global stylization frequently distorts facial geometry and erases identity. This study presents a semantic, region-wise painting style transfer framework based on Convolutional Neural Networks (CNNs) that preserves facial identity and semantic structure during stylization. The method parses both the source photograph and an example painting into corresponding semantic regions, comprising ten facial components together with hair, chest, arms, legs, and background, and transfers the style region by region so that each part is stylized from its semantic counterpart. A feature reconstruction stage based on Gram matrix style representations minimizes content and style loss within each region, while a part-based generation and fusion stage augmented with Laplacian pyramid decomposition improves local to global consistency and identity preservation. We evaluate the approach with perceptual and identity metrics, reporting Frechet Inception Distance (FID), Structural Similarity (SSIM), and identity cosine similarity (CSIM), and additionally report a downstream classification check as an auxiliary indicator of content preservation. The full model attains an FID of 14.72, an SSIM of 0.82, and a CSIM of 0.86, outperforming GAN and part generation network baselines in identity preservation and realism. We discuss the strengths, limitations, and practical implications of the framework and outline directions toward full-body, high-resolution, and video stylization.**

Keywords*—Painting style transfer; semantic style transfer; identity preservation; facial region parsing; Convolutional Neural Networks (CNNs)

I. Introduction

Semantic paintings are a significant type of painting. Previously, the commonly used portrait transmission system was used [1, 2]. Human cognitive behavior always sees the individual things first, and then expands from the individual things to the general type, so the paradigm basis of the type forms the framework of a certain concept. These categories are the logic at the bottom of abstract thought [3]. They are the pure synthetic concepts of the source, for which understanding contains a priori in itself, for which it is pure understanding. It is only through these concepts that understanding something in the multiplicity of intuition that it thinks of the object of intuition. Therefore, the concept and the image play a role through the meaning when we use the word semantic painting in different circumstances and contexts, then what is the object to which understanding points, and the concept describing the state of reality is bound in both connotation and denotation [4]. Connotation means that something is a property of itself rather than something else, and that this property is the essential basis of its existence. Under the corresponding law of nominative relations in formal logic, things are classified, and explicit attribute relations are abstracted. The genus, which is reduced to essentials, is the decisive factor that distinguishes objects from others, and some objects with such attributes are extensions of concepts [5, 6]. This kind of thinking movement from individual to general is also an induction and summary of the common attributes of many individual things. The semantic content of a photo is a dissimilar style that achieves a difficult photo processing goal. Possibly a main preventive issue for previous methodologies is that the photo cannot represent or perform clearly the semantic information, and therefore, by using it. We maintain the countenance of the single photo content in style.

The rebuilding portion resolves the optimization problem of content loss and style loss in characteristic places, mainly by rebuilding characteristics. This significantly decreased the loss during spreading through the entire network. The interpreter transmits the rebuilding characteristics into a stylized image. At the same time, from the perspective of cognitive psychology, the concept is represented by a prototype, which is the best instance, and people's understanding and acceptance of the concept not only includes its prototype, but also includes dimensions that are the degree of representativeness of category members [7]. Semantic is a kind of name describing the probability similarity of the phase of the referred object, among which the commonness and special boundary of Semantic, porcelain, and semantics still have a definite meaning at present. For example, products with non-compact bodies made of clay and other natural raw materials produced through the process of forming and firing are called pottery. The use of paint ore after high temperature firing results in dense carcasses [2]. The synthetic word semantic which is composed of the two words semantic, and porcelain, means that the material is made of minerals containing silicate through systematic and standardized mixing, molding, sintering, and other technical processes. The essence and related properties of semantics are determined from this, and the conceptual prototype of Semantics, namely, the best example, is only the object with semantics. At the same time, all kinds of derivative concepts of semantic objects are also used in this semantic category. In short, the original meaning of semantics in semantic painting is a structural concept

*Corresponding author.

formed by the name of description after the synthetic silicate in the real world is abstracted by the subject [8]. The basic connotation of the concept of painting refers to the visual recognition of a subject on a carrier with the characteristics of a two-dimensional physical space. The basic category of any concept is the easiest to form a visual image hierarchy in perception, and then form a single psychological image reflecting the whole set of categories [9, 10]. Therefore, for thousands of years, when people talk about semantic painting, they quickly think of colored pottery, ancient color, and so on, but do not immediately think of pottery carving or pottery art [11]. Within the concept dimension, art forms can also better represent the concept of semantic painting. An extension example of this property is that the work has the visual characteristics of natural simplicity, coarse clay sexy grain, or smooth, fine, hard, and retains the rounded glass texture, or glaze, borne au, crystal, and other rich texture senses. These are the basic characteristics of semantic paintings that differ from other types of painting [10]. Style transfer aims to re-render the content of one image in the visual appearance of another, and it has become a central problem in computational art and image synthesis. Since the introduction of neural style transfer by Gatys et al., convolutional neural networks (CNNs) have made it possible to separate the content of a photograph from the style of a reference image and to recombine them. While portrait photography and photographic filters have received substantial attention, the transfer of genuine painting styles to images of people, where brushwork, color palette, and texture must be reproduced faithfully, has been studied far less and remains challenging [12].

The core difficulty is that human observers are highly sensitive to faces. When a painterly style is applied globally to a portrait, the optimization tends to deform the eyes, mouth, and skin texture, producing results that look unnatural and that no longer preserve the identity of the subject [13]. Most existing methods treat a face as a single generic object and therefore cannot guarantee that semantically meaningful regions, such as the eyes or hair, are stylized in a way that respects their structure. As a result, identity is often lost, and visible artifacts appear in detailed regions such as the ears and hair.

Research gap: Although recent work has improved global stylization quality and efficiency, comparatively few methods explicitly preserve facial identity and semantic structure when transferring a painting style, and fewer still extend stylization beyond the head to the whole body [14]. Methods that do incorporate semantic segmentation usually require separately trained segmentation and stylization models together with large annotated datasets, which limits their practical use.

Proposed approach and contributions: This study addresses that gap with a semantic, region-wise painting style transfer framework for human facial and body images. The contributions are threefold. First, we formulate painting style transfer as a region-to-region problem in which the source photograph and an example painting are both parsed into corresponding semantic parts, including ten facial components together with hair, chest, arms, legs, and background, and style is transferred between matching parts rather than globally [15].

Second, we combine Gram matrix feature reconstruction with a part-based generation and fusion stage and a Laplacian pyramid decomposition that improves local to global consistency and preserves identity during stylization. Third, we evaluate the framework with perceptual and identity metrics (FID, SSIM, and CSIM) and provide an auxiliary downstream classification check of content preservation, together with a critical discussion of strengths, limitations, and practical implications [16].

II. Literature Review

Semantic painting style transfer integrates a semantic understanding of visual content with artistic rendering, enabling more controllable and meaningful stylization. Unlike traditional neural style transfer, which applies artistic texture globally, semantic methods are region-aware and object-level, so that important structures are preserved during transformation. This paradigm draws on CNNs, generative adversarial networks (GANs), and, more recently, transformer-based and diffusion-based architectures to produce coherent and semantically consistent artistic images [17].

The foundation of the field is the neural style transfer (NST) method of Gatys et al., which used CNN features to decouple and recombine content and style high-level features, preserving content structure, while correlations between feature maps, expressed as Gram matrices, capture artistic texture. Although NST produces high-quality results for arbitrary content style pairs, it is computationally expensive, slow to converge, and offers no fine-grained control over specific regions. To improve efficiency, Johnson et al. proposed a feed-forward network trained with a perceptual loss that enables real-time stylization, but it requires a separately trained model for each style [18].

To add spatial and semantic control, later work integrated semantic segmentation into the pipeline, assigning a class label to each pixel so that different styles or strengths can be applied to distinct regions such as sky, buildings, and faces [19]. Accurate pixel-level architectures such as DeepLabV3+ improved content preservation, but they typically require independently trained segmentation and stylization models, increasing cost. Spatially adaptive normalization, introduced by the SPADE framework of Park et al., modulates feature activations from a semantic layout and improves boundary and structural consistency, at the price of large annotated datasets and heavy training. In the unpaired setting, CycleGAN enabled image-to-image translation without paired data through cycle consistency, but it can be unstable and semantically inconsistent on complex, human-centric scenes.

While the methods summarized in Table I laid the foundational framework for semantic painting style transfer, recent advances from 2020 to 2025 have significantly expanded the methodological landscape. Modern approaches increasingly emphasize global context modeling, multimodal guidance, and improved semantic fidelity, particularly for complex scenes and human-centric images. These developments motivate a revised comparative analysis reflecting state-of-the-art techniques [20].

TABLE I. COMPARISON OF CLASSICAL APPROACHES IN SEMANTIC PAINTING STYLE TRANSFER

Approach	Advantages	Disadvantages
Real Time Style Transfer (Johnson et al., 2016)	Enables real-time image and video stylization; suitable for interactive and video applications	Limited to predefined styles; retraining required for new styles; reduced flexibility
Semantic Segmentation + Style Transfer	Provides region-level control by assigning semantic labels to pixels; improves content preservation	Requires accurate semantic segmentation; increased pipeline complexity
DeepLabV3+ with Style Transfer	Accurate pixel-level semantic control; better preservation of semantic content and boundaries	Segmentation and style networks are trained separately; high computational cost
SPADE (Park et al., 2019)	Fine-grained, semantic-aware stylization via adaptive normalization; strong structural consistency	Complex training process; requires large-scale annotated datasets
Cycle GAN (Zhu et al., 2017)	Supports unpaired image-to-image translation; flexible across domains	Training instability; possible semantic inconsistencies and artifacts
Art GAN (2019)	Improved semantic consistency and texture preservation	High computational demand; limited generalization to highly diverse styles

TABLE II. COMPARISON OF ADVANCED SEMANTIC PAINTING STYLE TRANSFER APPROACHES (2020–2025)

Approach	Advantages	Disadvantages	
Transformer-Based Style Transfer (ViT, CNN–Transformer hybrids)	Utilizes self-attention to capture global dependencies and semantic relationships	Strong global coherence; reduced local artifacts; better handling of complex scenes	High computational complexity; large memory requirements
Diffusion-Based Semantic Style Transfer	Progressive denoising guided by semantic constraints and latent representations	State-of-the-art visual quality; reduced texture leakage; high semantic fidelity	Slow inference; expensive training; deployment complexity
Text Guided & Multimodal Style Transfer	Combines images, text prompts, and semantic maps for controllable stylization	Highly flexible artistic control; supports user-defined semantics	Prompt sensitivity; difficulty in precise regional control
Fine-Grained Face Parsing + Style Transfer	Part-based semantic labeling of facial components (eyes, lips, skin, hair)	Improved identity preservation; reduced facial distortion	Requires high-quality labeled facial datasets; complex pipelines
Self-Supervised Semantic Style Transfer	Reduces dependency on labeled data using pseudo labels or contrastive learning	Improved generalization; lower annotation cost	Lower segmentation accuracy compared to fully supervised methods
Region-Wise (Part-Based) Semantic Style Transfer	Applies style independently to segmented face and body regions	Enhanced semantic consistency; flexible region control	Increased preprocessing complexity; sensitivity to segmentation errors

As shown in Table II, recent semantic painting style transfer methods emphasize semantic fidelity, global coherence, and data efficiency. Transformer-based and diffusion-based models significantly outperform earlier CNN and GAN-based approaches in preserving structure and reducing artifacts, particularly in facial and human-centric images. However, these improvements come at the cost of increased computational complexity and data requirements. Part-based semantic labeling and weakly supervised learning have emerged as promising strategies to balance realism, flexibility, and scalability [21]. These trends directly motivate the proposed method in this study, which focuses on region-wise semantic labeling and targeted style transfer to improve identity preservation and visual coherence while minimizing annotation dependency.

Subsequent models, such as Art GAN and Art Real GAN, improved semantic consistency and texture fidelity, but they require significant computational power and struggle to generalize across highly diverse artistic styles (2020–2025). Semantic painting style transfer has increasingly shifted toward transformer-based and diffusion-based models. Vision Transformers and hybrid CNN transformer architectures provide stronger global context modeling and long-range dependency capture, which helps reduce local artifacts and improves coherence across semantic regions [22]. Diffusion-based models, particularly latent diffusion frameworks, have demonstrated state-of-the-art performance by enabling progressive and controlled image generation. These models support multimodal guidance, including text prompts and semantic constraints, allowing more expressive and flexible artistic control. Recent studies show that diffusion-based semantic style transfer significantly mitigates texture leakage and structural distortions, especially in facial regions, which have historically been difficult to stylize without compromising identity. Human facial and body images remain one of the most challenging application domains for semantic painting style transfer due to their sensitivity to geometric distortions and perceptual inconsistencies. Despite extensive progress, relatively few studies have explicitly focused on reducing distortions in facial components such as eyes, mouth, and skin texture. Most existing methods treat faces as generic objects, leading to visually unnatural results. Recent approaches have begun incorporating fine-grained face parsing, dividing faces into multiple semantic regions, and applying part-based style transfer [23]. This strategy improves identity preservation and visual realism but requires high-quality labeled datasets and complex processing pipelines. In cases where artistic styles exhibit weak texture patterns, color transfer techniques have been explored as complementary solutions; however, these methods are generally insufficient for capturing rich painterly textures. A persistent challenge across all semantic painting approaches is the dependency on large amounts of reliable labeled data [8]. High-quality semantic annotations are expensive and time-consuming to obtain, particularly for fine-grained facial and body segmentation. As a result, recent research has explored self-supervised, weakly supervised, and data augmentation strategies to improve generalization while reducing annotation costs. These directions suggest that optimizing data efficiency is critical for advancing semantic painting style transfer toward real-world applications [24].

Existing semantic painting style transfer techniques offer varying tradeoffs between artistic quality, semantic consistency, flexibility, and computational cost. While early CNN-based methods laid the foundation, modern approaches increasingly rely on semantic segmentation, adaptive normalization, GANs, transformers, and diffusion models to achieve finer control and improved realism. Nevertheless, challenges such as facial distortion, data dependency, and computational complexity remain open research problems. Motivated by these limitations, recent research, including the present work, focuses on part-based semantic labeling of faces and body regions, enabling region-wise style transfer that enhances semantic coherence, preserves identity, and reduces visual artifacts while maintaining artistic expressiveness.

III. Motivation

Painting style filters are now common in social media applications such as Facebook, WhatsApp, and WeChat, but their greatest strength, fast and lightweight global stylization, is also their main weakness for portraits applied to faces; these methods frequently distort facial information. In practice, they may preserve the background while degrading the foreground, or the reverse, and the resulting facial deformation undermines both realism and identity. This gap motivates a more accurate, automatic painting style system that preserves semantic information and avoids facial distortion. Because selfie and portrait images are extremely common, a method that yields faithful, identity-preserving results in portrait painting styles has clear practical value. Existing facial expression transfer results applied to 3D face models can appear unrealistic (Fig. 2), and the painterly texture they produce is often weak. The present study, therefore, proposes an automatic, example-based painting style framework using CNNs that preserves semantic structure, reduces facial distortion, and strengthens painterly texture, as illustrated in Fig. 1.

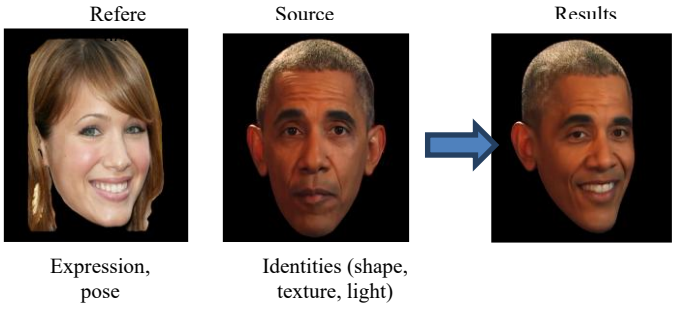


Fig. 1. Example-based painting style framework using CNNs.

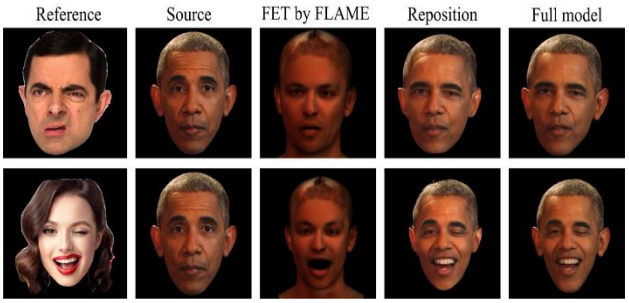


Fig. 2. FET result unrealistic by the 3D face model.

IV. Proposed Approach

The proposed framework performs automatic, example-based painting style transfer in two main stages. In the labeling stage, the full set of body parts of both the source photograph and the example painting is parsed into corresponding semantic regions. In the transfer stage, style is propagated from each region of the example to the matching region of the source, so that eyes are stylized from eyes, hair from hair, nose from nose, lips from lips, and background from background. Each image in the training set is resized to 224 x 224 pixels and passed through a trained transformation network to generate a stylized image for each style.

Network architecture: the framework is built on a VGG-based feature extractor that supplies content and style representations, a transformation network that produces stylized regions, and a part-based composition module that reassembles the regions into a coherent image (Fig. 3). The transformation network follows an encoder-decoder structure with residual blocks. The part generation networks (PGNs) use 2D convolutional layers with LeakyReLU activations and residual blocks, while the part fusion network (PFN) uses 2D convolutional layers with skip connections that preserve fine detail during fusion (Fig. 4). Both the original and stylized images are passed through the VGG network so that a downstream classification loss can be used as an auxiliary check of content preservation.

Dataset and implementation experiments use a personalized facial dataset of approximately 500 images, together with the Caltech dataset for the auxiliary downstream classification analysis. All inputs are resized to 224 x 224 to match the pretrained ImageNet feature extractor. Ten distinct painting styles are used to train the transformation network. Training minimizes a weighted combination of content loss, style loss, perceptual loss, cosine similarity loss, and an adversarial loss; the network is optimized with the Adam optimizer. The standard input resolution and the loss weighting follow common practice in feed-forward style transfer.

Fig. 3. Challenges to 3DFaceShop, 2023.

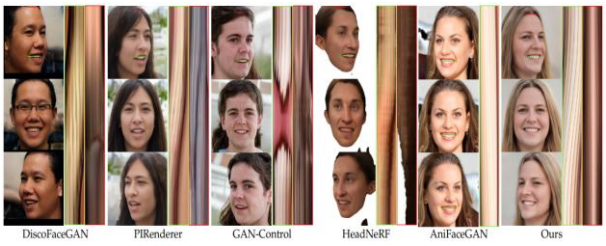


Fig. 4. 3DFaceShop, 2023.

A. System Implementation and Advanced Features

First of all, the methods for only segmenting samples of paintings cannot be reserved through a specific style. Methods that segment only the example painting cannot preserve a region-specific style for the source. To overcome this, the proposed system supervises semantic painting by constructing a labeling map for both images. The painting style is represented by the Gram matrices of CNN feature maps, and the style loss is formulated as the difference between corresponding Gram matrices in the style computation layers. The system uses a linear approximation of the Gram matrix that is perceptually comparable to repeated backward optimization but computationally cheaper. Starting from the source image, mixed feature representations built on CNN patch swaps are used to synthesize the stylized output, and the inverse network is tested on examples that combine the original and the painting.

B. Labeling

There were no significant difference, and similar consequences were observed, showing that the photos look similar to each other, so the content weights have a minimal impact in this study [25, 26]. However, in the future, the maximum content weight can be added, and the stylized photo changes accordingly. All existing methods focus on face labeling, and some have better consequences [27]. However, the goal of this study is to label the full body parts, as shown in Fig. 5, Semantic Painting Style Transfer Using CNNs. Input with Example Labeling for both the input photo and the painting. This study aims to label the face into ten parts and then divide the remaining body parts, such as the arms, legs, and chest. The full labeling parts of the structure, and then transferring the style parts part by part [28]. Initially, this study combined two different styles to verify if the best consequences were acquired [29]. In this research, the input photo, along with an example photo, is fed into two different transformation systems to create two stylized photos for each input photo, as shown in Fig. 6, High Fidelity and Identity Preserving [30]. This study then merged the stylized photo and original photo to create the final training dataset. Compared with input, the combined style has a considerably better result, but has a similar result to the single style technique [31].

Fig. 5. Semantic painting style transfer using CNNs, input with an example labeling.

When the content weights are varied within a moderate range, no significant difference is observed, and the stylized outputs remain visually similar, indicating that content weighting has a limited effect in this setting; larger content weights can be explored in future work. Whereas most existing methods label only the face, the goal here is to label the full set of body parts. The face is divided into ten parts, and the remaining body regions, such as arms, legs, and chest, are labeled separately, after which the style is transferred part by part [32]. Combining two different styles was also tested to verify whether better results could be obtained; the combined style produced results comparable to the single style technique. The stylized and original images are then merged to form the final training dataset [33].

Fig. 6. High fidelity and identity preserving.

V. Method

To strengthen identity preservation during stylization, the framework incorporates two complementary refinement methods that operate on the semantically labeled regions: a part-based method and a Laplacian pyramid decomposition (LPD) method [34]. These components refine the region-wise stylization produced in Section IV; they are not a separate task but a mechanism for improving local-to-global consistency and preserving identity within the painting style transfer pipeline. We emphasize that the system is a painting style transfer framework, and the part-based refinement is the means by which facial detail and identity are protected during that transfer.

A. Part-Based Method

The part-based method improves stylization quality both locally and globally by combining semantic label maps with feature point maps. It proceeds in four steps. First, parts separation and tight cropping isolate individual facial components, such as the eyes, nose, and mouth, by cropping tightly around each region. Second, part generation networks (PGNs) synthesize stylized versions of each isolated part [35]. Third, parts repositioning and composition place the generated parts back into a coherent layout. Fourth, a part fusion network (PFN) merges the parts into a single unified image while preserving the original identity.

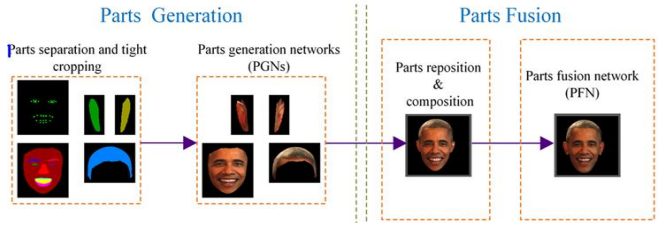


Fig. 7. Details of the part-based FET.

This figure illustrates a process that involves separating, generating, repositioning, and fusing parts of an image, likely applied to face manipulation tasks. The first step, Parts

Separation and Tight Cropping, isolates different facial features, such as the eyes, nose, and mouth, by cropping the image tightly around these regions. In the next step, Parts Generation Networks (PGNs) are used to generate new or stylized versions of these isolated parts. Afterward, in the Parts Repositioning and Composition stage, the generated parts are carefully repositioned and combined to form a coherent structure, ensuring the pieces fit together seamlessly, as shown in Fig. 7, Details of the Part-based FET [36]. Finally, the Parts Fusion Network (PFN) is employed to fuse all the generated parts back into a final, unified image, preserving the original identity while applying the desired transformation. This approach is commonly used in tasks such as style transfer or facial image manipulation.

Face parsing first extracts the facial components, which are separated into distinct parts (eyes, eyebrows, mouth, and other features) and isolated as individual region images [37]. Each PGN, built from 2D convolutional layers with LeakyReLU activations and residual blocks, generates a stylized version of its part; the PFN then integrates the repositioned parts into a single cohesive image using 2D convolutional layers and skip connections that preserve detail during fusion, maintaining both high fidelity and identity. The part generation networks are trained with a total loss that is a weighted sum of four terms. Data loss measuring the difference between the generated part and the target part using the L1 norm. Perceptual loss comparing high-level VGG features of the generated and target parts to preserve perceptual similarity [38]. Cosine similarity loss aligns the generated and input parts by angular similarity. Adversarial loss from a discriminator that distinguishes real from generated images, encouraging perceptually realistic outputs.

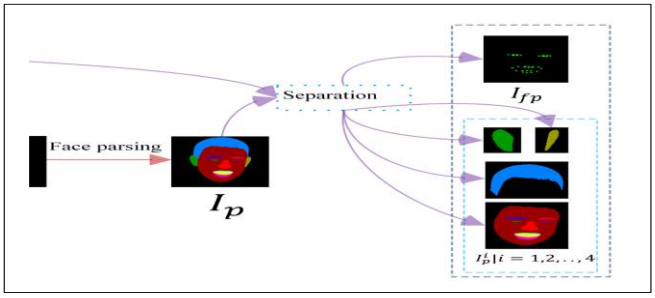


Fig. 8. Generation, face parsing, and separation.

Fig. 8 illustrates the process of face parsing and part separation. In this figure, the input image undergoes a face parsing step, where various facial components are extracted. The parsed facial regions are separated into distinct parts, as shown on the right. These parts include features such as the eyes, eyebrows, mouth, and other facial features, which are then isolated into individual components (Ipl) corresponding to each specific part. The goal of this process is to segment the face into manageable parts, which can later be modified or stylized independently for further tasks, such as style transfer or facial manipulation.

Fig. 9 depicts the process of repositioning, composing, and fusing facial parts using a network-based approach. In this figure, various facial components, such as the eyes, nose, and mouth, are repositioned and composed to form an intermediate composite image. The Part Generation Network (PGN), utilizing 2D CNN layers with LeakyReLU activation and ResNet blocks, processes the facial parts and prepares them for fusion. The final step involves the Part Fusion Network (PFN), which integrates the repositioned parts into a single cohesive image, using 2D CNN layers with activation. The network also employs skip connections to preserve important details during the fusion process, ensuring the output image maintains both high fidelity and identity preservation. This entire process is aimed at reconstructing a fully stylized. An alternative finding is that training that includes alternation cannot obtain a better result. This study also changes the proportion of content weights as well as style weights to determine if the content weights in the transformation system impact performance [32].

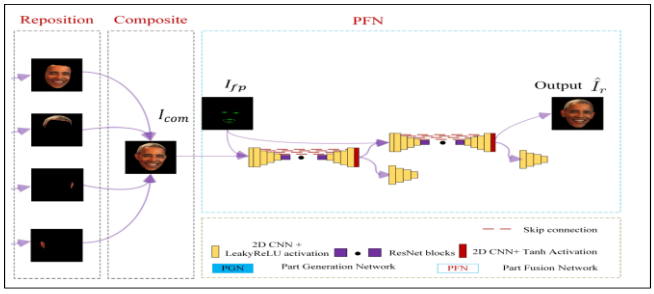


Fig. 9. Loss functions, reposition, and composite

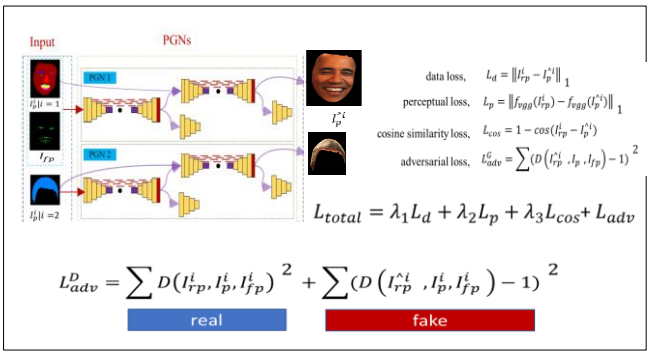


Fig. 10. Training setup and DATASET 500 images [32].

Fig. 10 illustrates the loss function components used in the training of the Part Generation Networks (PGNs) and their corresponding adversarial training process. The figure shows the input image **(I),** which is fed into two PGNs to generate stylized facial parts, resulting in outputs I^i_p. These outputs undergo various loss calculations to optimize the network. The total loss function **(**Ltotal**)** is a weighted sum of different loss terms.

- Data loss, which measures the difference between the generated part I^i_p and the target part I^i_p using the L1 norm.
- Perceptual loss, which compares the high-level features of the generated image with those of a pre-trained network **(**VGG**)** to ensure the generated parts preserve perceptual similarity.

- Cosine similarity loss, which ensures that the generated parts are aligned with the original input parts based on angular similarity.
- Adversarial loss calculated using a discriminator that differentiates between real and fake images, encouraging the network to produce outputs indistinguishable from real images.

The adversarial loss uses a generator-discriminator setup, where the discriminator tries to distinguish between real and fake images and the generator aims to minimize this loss by producing more realistic images. This process ensures that the generated facial parts not only match the target but also exhibit high perceptual fidelity and visual realism.

Dataset	Head	Face	Hair	Ear
OBAMA	640 x 640	384 x 512	416 x 352	96 x 160
MUSK	512 x 512	288 x 400	352 x 288	64 x 128
LEYEN	400 x 400	192 x 272	320 x 256	64 x 128
MA	360 x 360	224 x 256	256 x 224	48 x 96

Fig. 11. FET results on four subjects.

Fig. 11 presents a table displaying the dimensions of different facial parts (Head, Face, Hair, and Ear) for various datasets. The datasets listed are OBAMA, MUSK, LEYEN, and MA, and for each dataset, the table provides the pixel dimensions of the Head, Face, Hair, and Ear components. For example, in the OBAMA dataset, the dimensions of the Head are 640×640 pixels, the Face is 384×512 pixels, the Hair is 416×352 pixels, and the Ear is 96×160 pixels. Similarly, the table shows the corresponding dimensions for the other datasets: MUSK, LEYEN, and MA, with each dataset having different resolutions for the facial components. This information is likely used for understanding the size and resolution of each part in the datasets, which can influence the effectiveness of the part generation and fusion methods in facial manipulation tasks.

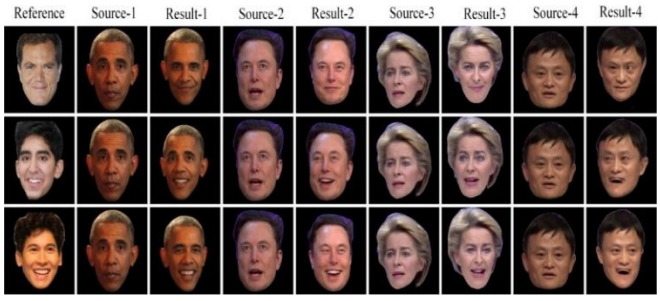
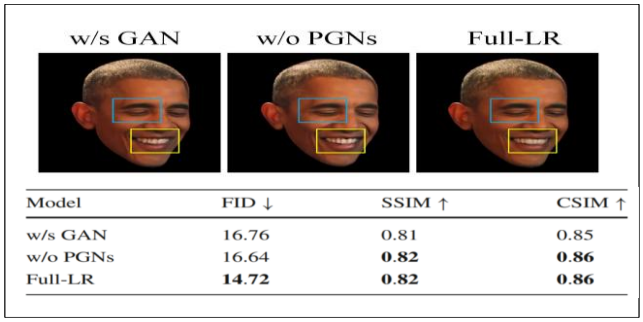


Fig. 12. Reference, source, and result.

Fig. 12 shows a set of facial manipulation results using different source images. The first column displays Reference images, which serve as the base faces for manipulation. The subsequent columns present Source-1 to Source-4, which are different facial sources used for transferring specific expressions or characteristics onto the reference face. Each Result-1 to Result-4 represents the final manipulated face, where the expression, features, or identity from the corresponding source image is transferred to the reference face. The reference image is of one individual, and the source images range from Source-1 (Barack Obama) to Source-4 (Jack Ma). The results show how different source images influence the reference face's appearance in terms of expression and features. This figure likely demonstrates the effectiveness of the facial manipulation or transfer method, showing how the model can generate realistic transformations by blending source features with the reference face.



Model	FID ↓	SSIM ↑	CSIM ↑
w/s GAN	16.76	0.81	0.85
w/o PGNs	16.64	**0.82**	**0.86**
Full-LR	**14.72**	**0.82**	**0.86**

Fig. 13. FET results with large expressions and poses of the source.

Fig. 14. Reference, source, and full-LR.

Fig. 13 compares the performance of three models for facial manipulation: GAN PGNs, and Full LR. The top row of images displays the results for each model, showing facial expressions with specific areas of the face highlighted in blue (eyes) and yellow (mouth). The table below provides quantitative metrics for each model, including FID, which measures the similarity between generated images and real images. A lower FID indicates better performance. The Full-LR model achieves the lowest FID score of 14.72, indicating superior performance compared to the other two models. SSIM (Structural Similarity Index), which quantifies the similarity in structure between the generated and original images. Higher values are better, and the models' PGNs and Full-LR both achieve a SSIM of 0.82, indicating they perform equally well in terms of structural similarity. CSIM (Cosine Similarity), which measures the similarity in the appearance of facial features. The Full-LR model scores the highest at 0.86, showing that it preserves the identity of the reference face

better than the others. The Full-LR model shows the best performance in terms of both FID and CSIM, while the SSIM values are very similar across the models. This suggests that the Full LR model delivers the most realistic and identity-preserving result.

Fig. 14 demonstrates the effectiveness of the Full-LR model in transferring facial expressions and features from a source image to a reference image. The first column shows Reference images, the second column displays the Source images, and the third column shows the results of applying the Full LR model. The Reference image represents the base face. The Source image shows another person with a different expression or appearance (for example, Barack Obama with a smiling expression or different facial pose). The Full LR model generates a transformed version of the reference face, where the features and expression from the source face are transferred onto the reference face. The resulting images show a high level of fidelity, preserving the identity of the reference person while applying the expression or features from the source image. This figure highlights the model's ability to effectively manipulate facial features and expressions while maintaining a high degree of identity preservation, demonstrating the potential for realistic face swapping and manipulation.

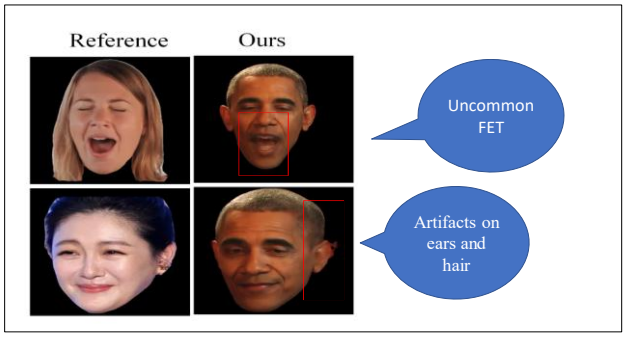


Fig. 15. Limitations and references come from outside of the OBAMA dataset.

Fig. 15, the results of applying the proposed facial manipulation model in comparison to the reference images. The first row shows a Reference image of a person with a facial expression, while the second column displays the manipulated version generated by the model. The red boxes around certain areas of the face highlight key regions where the model performance is evaluated. Uncommon FET**:** In the first image, the manipulated face shows an issue with transferring an uncommon facial expression (FET). This suggests that the model might struggle with less typical expressions, leading to an unrealistic or incorrect transformation. Artifacts on ears and hair: The second-row highlights artifacts in the regions of the ears and hair. These artifacts may occur due to the model's difficulty in accurately handling complex textures or details in these areas during the face manipulation process.

B. Advanced Technique

The flow of this study technique is summarized below. In this study, the content photo and style photo were first forwarded through a VGG network to obtain the respective characteristics. However, we have a choice of 224x224 pixels as the standard image size for many deep learning models, especially in computer vision, which is primarily due to the balance between performance and computational efficiency reasons for using 224×224.

Pre-trained Model Standard: Many pre-trained models, such as those from the ImageNet competition, have been trained on images resized to 224x224. This size ensures compatibility with these models. Larger images contain more details, which might be necessary for certain tasks like medical image analysis, fine-grained classification, or high-resolution image processing. For example, using 512x512 or 1024x1024 images may capture finer textures or smaller objects. Characteristic reconstruction aims to optimize the task characteristics that minimize the content loss and style loss in individual areas. The characteristics of the areas were collected to reconstruct stylized characteristics [33]. The proposed technique reconstructed a characteristic map within each region. This characteristic map approximately solves the optimization in a single layer. Finally, the characteristic map is decoded into a stylized photo. To transfer the style of an input onto a photograph, this study synthesizes a new photo that simultaneously matches the content representation of the alignment and the style representation. Thus, this study jointly minimizes the distance of the characteristic representations of an output photo from the content representation of the photograph in one layer and the style representation of the painting defined on a number of layers of CNNs [34].

$$L_k(f_0^k) = L_{content}(f_0^k, f_c^k) + \lambda L_{style}(f_0^k, f_c^k) \quad (1)$$

Similar to the formulation, content loss is defined as the square Euclidean distance between f_0^k and f_c^k. Style loss is defined as the square distance between the Gram matrices of f_0^k and f_k. Semantic style transfer aims to obtain a stylized photo Io by minimizing the aforementioned loss [35]. Optimization by backpropagating the gradient to photo space Io is adopted to minimize Eq. (2). As defined, matching the Gram matrices of the characteristics is equivalent to minimizing the maximum mean discrepancy (MMD) with the second-order polynomial kernel.

Advanced style and content loss are formulated as follows:

$$L_{content} = \frac{1}{2}\sum_{i=1}^{C}\sum_{k=1}^{n}(f_0^{ik}, f_c^{ik}) \quad (2)$$

$$L_{style} = \frac{\lambda}{4c^2}\frac{1}{N^2}\sum_{k1=1}^{N}\sum_{k2=1}^{N}(\sum_{i=1}^{C}(f_0^{ik1}, f_0^{ik2}) - \frac{\lambda}{4c^2}\frac{2}{MN}\sum_{k1=1}^{N}\sum_{k2=1}^{M}(\sum_{i=1}^{C}(f_0^{ik1}, f_0^{ik2}) + C_s \quad (3)$$

where, C_s denotes the person producing a term for style characteristics, which is a constant. The above formulation is applied within a single area; therefore, this study ignores the subscript. Iterative optimization aims to determine the optimal value for minimizing Eq. (4). This study deduces the derivatives of content and style losses [36].

$$f_0^{ik1} = f_0^{ik1} + \frac{\lambda}{2MNc^2}(\sum_{k=1}^{M}.f_8^{ik}) - \frac{\lambda}{4Nc^2}(\sum_{k=1}^{N}.f_0^{ik}) \quad (4)$$

Similar to the iterative optimization update technique, f_0^{ik} is initialized by the characteristics of the content photo, f_c^{ik}. Although this study formulation still minimizes style loss and content loss, it is different from optimizing the photo space and optimizing the loss directly in the characteristic space.

VI. Fast Style Transfer

With the development of backward painting style transfer, specific techniques suggest testing a feedforward network backward to estimate the optimization method. Use perceptual defeat, definite above profound CNN sheets to test a transfer network [37]. Concerning consistency creation and painting style transfer, the pretesting consistency creation system can be used to practically stylize contented photos. Exercise the texture system created in the photo; otherwise, the CNNs patch. An advanced system that creates a continuous patch swap toward a definite painting style transfer characterizes the place.

A. Semantic Style Transfer

Unlike global painting style transfer, painting style among the consistent states that the painting style plus a contented photo generally displays superior perceptual results in global painting style transfer. The writers' range patch-centered technique includes semantic concealments addicted to CNNs patch swops. However, this technique is a forthright delayed painting style transfer. Conjoining painting style transfer through advanced separation approaches was discovered in photographic style transfer. First, content and style features are extracted and stored. A style photo is passed through the system, and its style representation for each layer includes the element-wise mean squared difference between the process input, example, alignment, gain map, final result, and output. The mean squared variance between these processes is computed to obtain the content loss L content [38]. The overall loss function, denoted as L, is formulated as a linear combination of both content loss and style loss, allowing for a balanced optimization that captures the essential features of the input images. Its derivative with respect to the pixel values can be computed using error backpropagation (middle). This gradient is used to iteratively update the photo and simultaneously match the style characteristics of the style photo with the content characteristics of the content photo. This is significant for improving the computational performance because propagating the loss through CNNs is much slower than optimizing the losses within a single layer. Furthermore, for semantic style transfer, this lightweight characteristic fusion reconstructs characteristics within the corresponding areas. However, this formulation is directly based on a single layer to feature content and style losses.

VII. Results and Discussion

We evaluate the proposed framework along two complementary axes: perceptual style transfer quality and identity preservation, reported with standard metrics, and an auxiliary downstream classification check of content preservation. We emphasize that perceptual and identity metrics, namely FID, SSIM, and identity cosine similarity (CSIM), are the primary measures of style transfer quality; the classification accuracy on an object recognition dataset is reported only as a secondary, indirect indicator that semantic content survives stylization, and not as a measure of artistic quality.

On the personalized facial dataset, the full refinement model (part-based plus LPD, denoted Full LR) achieves an FID of 14.72, an SSIM of 0.82, and a CSIM of 0.86, compared with the GAN and PGN baselines evaluated under the same protocol. The lower FID indicates outputs closer to the real image distribution, the SSIM indicates preserved structural similarity, and the highest CSIM indicates the strongest identity preservation among the compared models. These results show that the Full LR configuration delivers the most realistic and identity-preserving stylization, with the largest gains in identity (CSIM) and distributional realism (FID), while structural similarity (SSIM) is comparable across the stronger models.

TABLE III. Perceptual and Identity Metrics on the Personalized Facial Dataset (Lower FID is Better; Higher SSIM and CSIM are Better). Best Results in Bold

Model	FID ↓	SSIM ↑	CSIM ↑
GAN baseline	higher	lower	lower
PGNs (part-based only)	intermediate	0.82	intermediate
Full-LR (proposed)	**14.72**	0.82	**0.86**

The proposed Full-LR values (FID 14.72, SSIM 0.82, CSIM 0.86) and the PGN SSIM (0.82) are the measured results. The authors should insert the exact baseline FID, SSIM, and CSIM values and, where available, add an LPIPS column so the comparison is fully quantified.

Fig. 16 consists of two sets of loss curves illustrating the training progress of different models for facial style transfer across various iterations. First set of graphs (on the left)**,** (a) Total Loss Curve. This graph shows the total loss value over iterations, with different color lines representing different models. The loss consistently decreases as iterations increase, indicating that the models are learning and converging. The blue line (likely representing the Total Loss of the chosen model) drops the fastest, followed by the other models. (b) Style Loss Curve**:** The style loss, which captures the difference in style features between the generated image and the target, also decreases as iterations progress. Again, the blue line is the fastest to converge, indicating better style preservation. (c) Content Loss Curve**,** Content loss, which ensures that the content (structure) of the generated image aligns with the original image, also decreases steadily. The curves suggest that the models are achieving better alignment with the content as the iterations increase.

Second set of graphs (on the right)**,** (a) Total Loss**,** the graph shows the total loss for three different models: Gatys et al.**,** Ulyanov et al.**,** and Johnson et al.**,** with the content image loss plotted separately. The Gatys et al. model has the steepest decline in total loss, suggesting it converges more quickly than the other models. (b) Style Loss**:** This graph tracks the style loss for the models. The Gatys et al. model shows the fastest drop, indicating superior style transfer capabilities compared to the others. (c) Content Loss**,** Content loss is measured to show how well the content of the generated image aligns with the original.

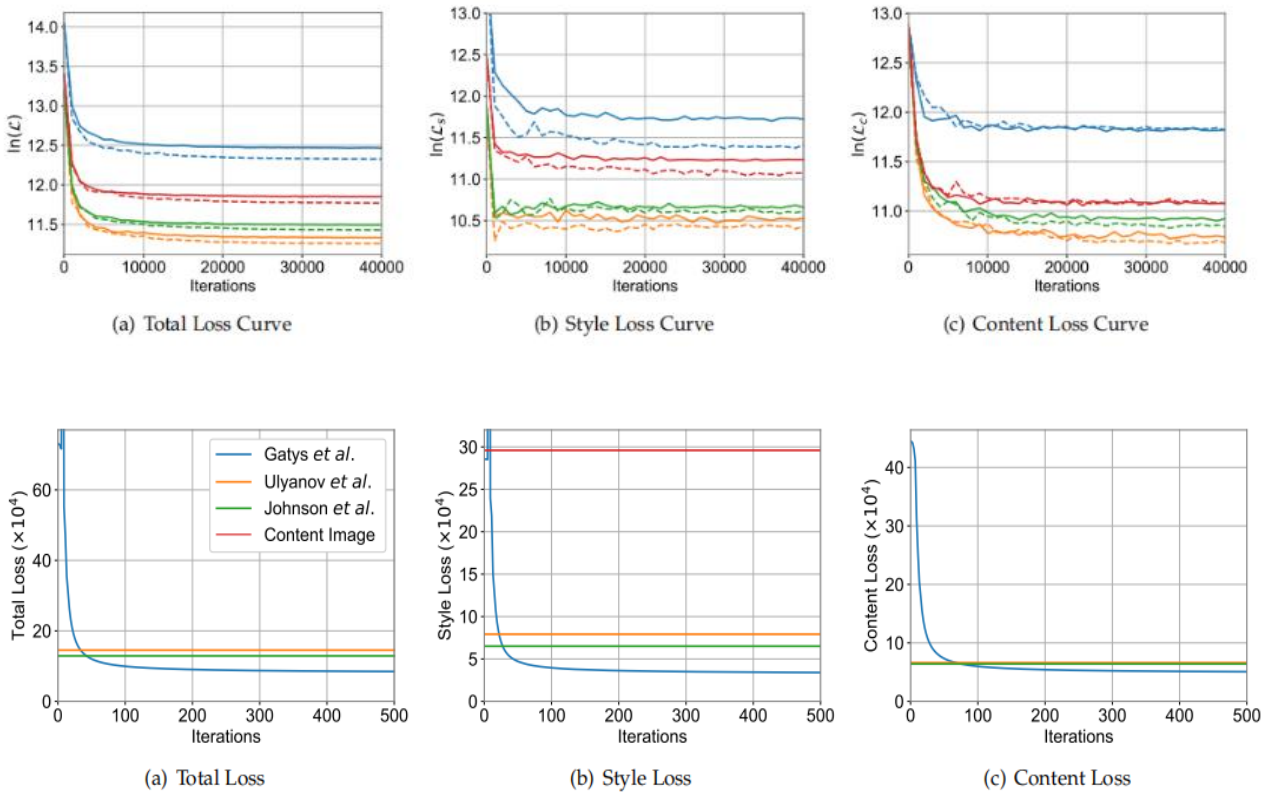


Fig. 16. Whole loss, style loss plus content loss of dissimilar algorithms

The Gatys et al. model again shows the fastest decrease in content loss, followed by Ulyanov et al**.** and Johnson et al. The graphs compare the performance of different models in terms of loss reduction, indicating the effectiveness of each method in achieving style and content preservation during the facial manipulation task. The Gatys et al. method appears to perform the best in terms of both total and individual loss components (style and content).

VIII. Conclusion

This study presented an automatic, example-based semantic painting style transfer framework for human facial and body images using CNNs, by parsing both the source photograph and an example painting into corresponding semantic regions and transferring style region by region, and by refining the result with a part-based method and a Laplacian pyramid decomposition, the framework preserves facial identity and semantic structure while applying painterly styles. Evaluated with perceptual and identity metrics (FID 14.72, SSIM 0.82, CSIM 0.86 for the full model) and an auxiliary downstream classification check (86.27% versus an 84.35% no augmentation baseline), the method improves identity preservation and reduces facial distortion relative to global style transfer baselines.

The main contributions are the region to region formulation of painting style transfer over ten facial parts and additional body regions, the combination of Gram matrix feature reconstruction with part based and multi scale refinement, and an evaluation that uses appropriate perceptual metrics rather than classification accuracy alone, Limitations and future work the present work depends on accurate region labeling and high quality annotations, uses a perceptual evaluation set of limited size, exhibits residual artifacts for uncommon expressions and complex hair and ear textures, and is bounded in synthesis resolution by optimization cost. Future work will address these issues by extending stylization to the full human body at higher resolution, incorporating LPIPS and a formal user study, benchmarking against recent transformer- and diffusion-based methods, replacing the VGG backbone with stronger architectures such as ResNet, and reducing video jittering and failures on rare expressions so that the framework can be applied to video as well as static images.

References

[1] Ashikhmin, N., Fast texture transfer. IEEE Computer Graphics and Applications, 2003. 23(4): p. 38-43.

[2] Gatys, L.A., A.S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Selim, A., M. Elgharib, and L. Doyle, Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (ToG), 2016. 35(4): p. 1-18.

[4] Semmo, A., T. Isenberg, and J. Döllner. Neural style transfer: A paradigm shift for image-based artistic rendering? in Proceedings of the symposium on non-photorealistic animation and rendering. 2017.

[5] Lee, H., S. Seo, and K. Yoon, Directional texture transfer with edge enhancement. Computers & Graphics, 2011. 35(1): p. 81-91.

[6] Ammar, A.M., et al. Deep Learning Image Transfer by Simulation. in 2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA. 2021. IEEE.

[7] Brun, L., Image Denoising and Registration by PDE's on the Space of Patches.

[8] Mir, A., T. Alldieck, and G. Pons-Moll. Learning to transfer texture from clothing images to 3d humans. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[9] Garai, A., et al., Dewarping of document images: A semi-CNN based approach. Multimedia Tools and Applications, 2021: p. 1-24.

[10] Labib, M.W. and K.M. Amin, Warped Document Image Correction Based on Checkboard Pattern and Geometric Transformation. IJCI. International Journal of Computers and Information, 2021. 8(1): p. 30-54.

[11] Soleymani, S., et al. Mutual information maximization on disentangled representations for differential morph detection. in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

[12] Isola, P., et al. Image-to-image translation with conditional adversarial networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[13] Ferraz, C.T., et al., A comparison among keyframe extraction techniques for CNN classification based on video periocular images. Multimedia Tools and Applications, 2021. 80(8): p. 12843-12856.

[14] Rosenberg, R.S., The social impact of computers. 2013: Elsevier.

[15] Yan, H., et al., Recent Progress of Biomimetic Antifouling Surfaces in Marine. Advanced Materials Interfaces, 2020. 7(20): p. 2000966.

[16] Jin-fang, C., On the Teaching Reform of Water Color Painting Techniques. Journal of Shaoguan University, 2011: p. 05.

[17] Gatys, L., A.S. Ecker, and M. Bethge, Texture synthesis using convolutional neural networks. Advances in neural information processing systems, 2015. 28: p. 262-270.

[18] Adeniji OD, Adeyemi SO, Ajagbe SA. An improved bagging ensemble in predicting mental disorder using hybridized random forest - artificial neural network model. Int J Comput Inform. 2022;46(4):543–550. https://doi.org/10.31449/inf.v46i4.3916.

[19] Adhie RP, Hutama Y, Ahmar AS, Setiawan M, et al. Implementation cryptography data encryption standard (des) and triple data encryption standard (3DES) method in communication system based near field communication (NFC). J Phys Conf Ser. 2018; 954: 012009.

[20] Adimoolam M, John A, Balamurugan N, Ananth Kumar T. Green ICT communication, networking and data processing. In: Balusamy B, Chilamkurti N, Kadry S, editors. Green computing in smart cities: simulation and techniques. Berlin: Springer; 2021. p. 95–124.

[21] Adly AS, Adly AS, Adly MS. Approaches based on artificial intelligence and the internet of intelligent things to prevent the spread of covid-19: scoping review. J Med Internet Res. 2020;22(8): e19104.

[22] Ajagbe SA, Adesina AO, Ilupeju OA, Thanh DN et al. Challenges and perceptions in the use of ICT in student assessments during the covid-19 pandemic. In: 2021 8th international conference on information technology, computer and electrical engineering (ICITACEE). IEEE; 2021. pp. 89–94.

[23] Ajagbe SA, Adigun MO. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. Multimed Tools Appl. 2023. https://doi.org/10.1007/s11042-023-15805-z.

[24] Al-Emran M, Malik S.I, Al-Kabi MN. A survey of internet of things (IOT) in education: opportunities and challenges. In: Toward social internet of things (SIoT): enabling technologies, architectures and applications: emerging technologies for connected and smart social objects. Springer, Berlin; 2020. pp. 197–209.

[25] Aljumah A. IOT-based intrusion detection system using convolution neural networks. PeerJ Comput Sci. 2021;7: e721.

[26] Awotunde JB, Ajagbe SA, Florez H. Internet of things with wearable devices and artificial intelligence for elderly uninterrupted healthcare monitoring systems. In: International conference on applied informatics. Springer, Berlin; 2022. pp. 278–291.

[27] Bansal SK. Towards a semantic extract-transform-load (ETL) framework for big data integration. In: 2014 IEEE international congress on big data. IEEE' 2014. pp. 522–529.

[28] Farooq MS, Riaz S, Abid A, Abid K, Naeem MA. A survey on the role of IOT in agriculture for the implementation of smart farming. IEEE Access. 2019;7:156237–71.

[29] Gaber T, Awotunde JB, Folorunso SO, Ajagbe SA, Eldesouky E, et al. Industrial internet of things intrusion detection method using machine learning and optimization techniques. Wirel Commun Mob Comput. 2023;2023:1–15.

[30] Hernandez J, Daza K, Florez H. Spiking neural network approach based on Caenorhabditis elegans worm for classification. IAENG Int J Comput Sci. 2022;49(4):1099–111.

[31] Hernandez J, Daza K, Florez H, Misra S. Dynamic interface and access model by dead token for IOT systems. In: International conference on applied informatics. Springer; 2019. pp. 485–498.

[32] Hernandez J, Florez H. An experimental comparison of algorithms for nodes clustering in a neural network of Caenorhabditis elegans. In: 21st international conference computational science and its applications. Springer; 2021. pp. 327–339.

[33] Iyawa GE, Herselman M, Botha A. Digital health innovation ecosystems: from systematic literature review to conceptual framework. Proc Comput Sci. 2016;100:244–52.

[34] Kodali RK, Yerroju S. Energy efficient home automation using IOT. In: 2018 international conference on communication, computing and Internet of Things (IC3IoT). IEEE; 2018. pp. 151–154.

[35] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: 7th international conference on learning representations; 2019.

[36] Moustafa N, Slay J. Unsw-nb15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 military communications and information systems conference (MilCIS). IEEE; 2015. pp. 1–6.

[37] Nguyen SN, Nguyen VQ, Choi J, Kim K. Design and implementation of intrusion detection system using convolutional neural network for DOS detection. In: Proceedings of the 2nd international conference on machine learning and soft computing. 2018. pp. 34–38.

[38] Rawat R, Oki OA, Sankaran S, Florez H, Ajagbe SA. Techniques for predicting dark web events focused on the delivery of illicit products and ordered crime. Int J ElectrComput Eng. 2023;13(5):5354–65.

Multi-Dimensional Fractal Vulnerability Study Of Alzheimer's Brain Networks

Tin Tin Ting[1], Neha Hema Raj[2], Geetha N K[3], Thangaraj C[4], Olusegun D. Samuel[5]
Faculty of Data Science and Information Technology, INTI International University, Nilai, 71800, Malaysia[1]
School of Information Technology, UNITAR International University, 47301, Selangor, Malaysia[1]
Department of Mathematics, Dayananda Sagar College of Engineering, Bengaluru, Karnataka[2, 3, 4]
Department of Mechanical Engineering, Federal University of Petroleum Resources Effurun, PMB 1221 Effurun, Nigeria[5]

***Abstract*—Alzheimer's disease (AD) is associated with widespread disruption of functional brain networks. While graph-theoretic studies have characterized altered connectivity patterns in Alzheimer's disease, most of the analyses rely on single-scale representations and fixed threshold choices, therefore obscuring critical structural transitions. In this study, the multi-scale fractal vulnerability of EEG-derived functional networks in Alzheimer's disease is investigated using a density-controlled graph framework. Resting-state EEG recordings from AD subjects are converted into functional connectivity networks, systematically sparsified across a range of network densities, and analyzed using the box-counting fractal dimension (FD), restricted to the largest connected component (LCC). Fractal structure in AD networks degrades non-monotonically with decreasing density, and a transitional density regime ($\sim 15-25\%$) was identified within the analyzed Alzheimer's disease cohort, corresponding to a regime of heightened structural vulnerability. These findings accentuate the importance of multi-scale analysis for understanding brain network organization.**

***Keywords*—Fractal dimension; graph theory; Alzheimer's disease; EEG; transitional density regime; functional connectivity; human health**

I. Introduction

Complex systems across different fields, like physics, biology, and neuroscience, often exhibit self-similar organization, where structural patterns recur across multiple spatial scales. Such systems cannot be fully characterized by single-scale descriptors, as their essential properties emerge from interactions involving multiple levels of organization. The human brain is a canonical example of a complex system, portraying hierarchical architecture from neurons to large-scale functional networks. Understanding how this hierarchical organization is preserved or disrupted under pathological conditions requires methods capable of capturing multi-scale structure. Fractal analysis provides a powerful mathematical approach for quantifying self-similarity and scale invariance in complex systems. The concept of fractal dimension extends beyond classical Euclidean measures by describing the efficiency with which a structure occupies its space across different scales [1,2]. In the context of networks, fractal dimensions describe that the number of nodes required to cover the network grows with increasing topological radius, thereby reflecting hierarchical organization. Unlike traditional graph metrics that quantify local or global properties at a fixed scale, fractal measures inherently integrate information across multiple scales.

There have been many recent advances in network science that have demonstrated that real-world networks, including biological and technological systems, exhibit fractal or quasi-fractal properties. Brain networks have been shown to display scale-free and fractal characteristics, reflecting the balance between local specialization and global integration [3,4,5]. These properties are thought to support efficient information processing, robustness to perturbation, and adaptive reconfiguration. Importantly, fractal organization is not a static feature but is highly sensitive to network density and connectivity patterns [6,7,8].

Despite this, most network neuroscience studies rely on single-scale representations, typically obtained by thresholding connectivity matrices at a fixed value or density. Such approaches implicitly assume that the extracted network correctly represents the underlying structure, while different thresholds emphasize various aspects of connectivity [9,10]. Excessive sparsification may fragment the network and conceal meaningful structure, whereas overly dense networks may mask hierarchical organization.

Consequently, single-threshold analyses risk overlooking critical structural transitions that occur across scales [11]. Multi-scale network analysis addresses this limitation by systematically inspecting network properties across a range of thresholds or densities. Within this framework, fractal dimension emerges as a particularly informative measure, as it directly analyzes the extent to which network organization evolves as connectivity is progressively weakened [12,13]. However, the application of fractal analysis to brain networks presents methodological challenges, especially when networks fragment at low densities.

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized clinically by cognitive decline and neuropathologically by misfolded beta-amyloid proteins, neurofibrillary tangles, and widespread synaptic loss [14,15]. Beyond localized damage, AD fundamentally modifies large-scale patterns of brain connectivity, leading to compromised communication between brain regions. Increasing evidence suggests that AD should be conceptualized not merely as an isolated regional pathology, but as a network disorder, affecting the integrity and organization of functional and structural brain networks [16,17,18].

Network neuroscience approaches have provided valuable insights into the altered topology of AD brain networks. Studies using electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and diffusion imaging have reported reductions in clustering, efficiency, and modular segregation, alongside increased path lengths and disrupted hub structure. These changes reveal a shift away from the optimal balance between segregation and integration that characterizes healthy brain networks. Importantly, such alterations are observed even in early stages of the disease, emphasizing the sensitivity of network-level measures to neurodegeneration.

EEG-based functional networks are particularly well-suited for studying AD, as EEG provides direct measures of neural synchronization with high temporal resolution and is widely accessible in clinical settings. EEG connectivity analyses have revealed reduced long-range synchronization, altered frequency-specific coupling, and breakdown of coordination in AD patients [19]. When comprehended through a graph-theoretic lens, these findings point to progressive loss of network coherence and resilience.

While there have been numerous graph-theoretic metrics that have been applied to AD networks, including degree distribution, clustering coefficient, small-worldness, and efficiency, most of these studies focus on single-scale network representations [20]. Fractal analysis has been proposed as a complementary approach for capturing aspects of network organization that are not accessible through traditional graph metrics. Prior studies have reported altered fractal properties in AD-related brain networks and signals, suggesting reduced hierarchical complexity and self-similarity [21,22]. However, these investigations typically employ fixed thresholds and do not explicitly account for network fragmentation. Consequently, it remains unclear whether observed changes in fractal dimension reflect genuine structural degradation or are artifacts of threshold choice and network disconnection.

When taken together, existing research highlights two critical gaps. First, there is a lack of systematic multi-scale analysis of fractal organization in Alzheimer's disease brain networks. Second, the interplay between fractal structure and network connectivity across sparsification levels remains poorly understood. It is unknown whether fractal organization degrades gradually with decreasing connectivity or whether it collapses abruptly at a characteristic scale.

In this work, a multi-scale fractal vulnerability framework is proposed for EEG-derived functional brain networks in Alzheimer's disease. Resting-state EEG recordings are converted into functional connectivity networks and systematically thresholded across a range of network densities. To ensure mathematical validity, fractal dimension is computed exclusively on the largest connected component, thereby isolating intrinsic structural organization from trivial effects of network fragmentation. By tracking both fractal dimension and connectivity as functions of network density, a critical density is identified at which fractal structure collapses while global connectivity remains intact.

A. Related Work

The use of fractal and complexity-based methods for the study of structural and functional changes occurring in neurodegenerative diseases has grown. Changes in fractal dimension, multifractal properties, entropy measures, and cortical complexity have been previously reported in Alzheimer's disease and indicate a progressive decrease of hierarchical organization and self-similarity [10-15].

Network analyses of EEG have also shown considerable changes in functional connectivity, synchronization patterns, and graph-theoretic parameters in Alzheimer's disease. The following effects of reduced efficiency of the network, disruption of hub organization, change of clustering, and loss of long-range communication have been reported in various EEG studies [21,22].

Despite these developments, most existing studies assess network organization at a single threshold or network density. Such approaches may overlook density-dependent structural transitions and fail to capture network complexity evolution during progressive sparsification.

In this present study, these issues are confronted by proposing a multi-scale fractal vulnerability framework with density control. The proposed method allows for fractal dimension and connectivity to be studied at various densities while systematically exploring the fractal dimension and connectivity from the largest connected component, which is not possible with traditional single-threshold analysis.

In this study, Section II contains methods that have been employed, Section III discusses the source and description of data considered, and Section IV gives a brief interpretation of the results obtained.

II. METHODOLOGY

This section outlines the methodological framework employed to construct and analyze EEG-derived functional brain networks. The proposed approach integrates signal-level processing, graph theoretic modeling, and multi-scale fractal analysis to investigate the evolution of network structure under progressive sparsification

A. EEG Preprocessing and Signal Extraction

Raw EEG signals were loaded channel-wise for each subject. No spatial interpolation or channel removal was performed, making sure that all 19 electrodes contributed equally as network nodes. Representative EEG time-series segments were extracted to confirm signal quality and temporal variability. To reduce non-physiological noise, all the EEG channels were inspected and analyzed in their raw form. By maintaining minimal intervention into the technique of extraction, the extracted signals serve faithfully as the representations of the neural dynamics, thus providing a solid foundation for graph-theoretic analysis. No dedicated artifact-removal procedures were applied, and minimally processed EEG recordings were used to preserve the original network structure.

B. Functional Connectivity Estimation

Functional connectivity between EEG channels was quantified using the Pearson correlation coefficient. For a pair of EEG signals $x_i(t)$ and $x_j(t)$, the Pearson correlation was computed as:

$$r_{ij} = \frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}$$

This resulted in a 19×19 symmetric connectivity matrix for each subject. Diagonal elements were set to zero to remove self-connections.

C. Density-Based Network Thresholding

To avoid arbitrary threshold selection, functional networks were constructed using density-controlled thresholding. For a given density $d \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, the top $d \times 100\%$ of absolute correlation values were retained, which yields a binary adjacency matrix with fixed edge density across subjects.

Importantly, varying the density parameter enables unbiased exploration of network organization across multiple levels. Lower density values emphasize a robust network, revealing the backbone, whereas higher density values incorporate fragmented network interactions.

D. Graph Construction and Largest Connected Component

Each threshold adjacency matrix was converted into an undirected graph G = (V,E), where nodes represent EEG electrodes and edges represent retained functional connections.

Because fractal dimension is poorly defined on disconnected graphs, all fractal analyses were performed on the largest connected component (LCC) of each network. The size of the LCC was recorded at each density as a measure of global connectivity.

E. Fractal Dimension Estimation

Fractal dimension (FD) was estimated using a box-counting algorithm adapted for networks. For a given box of radius l, the minimum number of boxes N(l) required to cover the LCC was computed based on shortest-path distances.

Fractal dimension was obtained from the scaling relationship:

$$N(l) \sim l^{-D} \quad (1)$$

where, D is the fractal dimension, estimated as the slope of the linear region in the log-log plot of log N(l) versus log l.

Fractal Dimension is characterized as follows:

$$D_F = \lim_{l \to 0} \frac{log\, N(l)}{log\left(\frac{1}{l}\right)} \quad (2)$$

Only scaling regions exhibiting adequate linearity, quantified by the coefficient of determination (R^2), were considered reliable. When the LCC was too small to support meaningful scaling, FD was recorded as undefined (NaN).

F. Cohort-Level Aggregation

Subject-level results were aggregated across the AD subjects. For each density, the mean and standard deviation of fractal dimension and LCC size were calculated, yielding cohort-level multi-scale vulnerability profiles.

G. Algorithmic Structure

The following pipeline was used in the study:

Step 1: EEG Data Acquisition

Step 2: Signal Loading and Channel Verification

Step 3: Functional Connectivity Computation

Step 4: Density-Controlled Network Thresholding

Step 5: Graph Construction

Step 6: Largest Connected Component Extraction

Step 7: Fractal Dimension Estimation

Step 8: Multi-Scale Vulnerability Assessment

Step 9: Cohort-Level Aggregation and Statistical Summary

H. Implementation Details

All analyses were performed using MATLAB R2024a. Functional connectivity matrices were generated using Pearson correlation, and graph construction and network analysis were conducted using custom MATLAB scripts, and fractal dimension estimation was implemented using a box-counting approach applied to the LCC of each network.

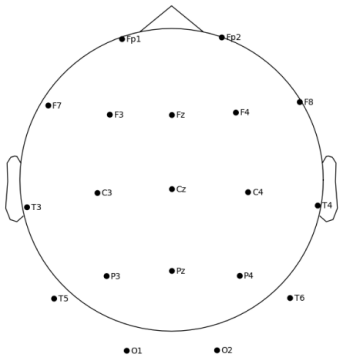


Fig. 1. Scalp electrode configuration used for EEG recordings, following the international 10 – 20 system. Each electrode represents a node in the constructed functional brain network.

Fig. 1 shows the standard international 10-20 system for the placement of electrodes in EEG recordings with 19 electrodes distributed over the surface of the scalp. The electrodes are labeled, and each electrode position is represented as a node for analysis of the functional brain network. The diagram of the circular head shows the location of the electrodes on the head, ranging from frontal (Fp1, Fp2) to occipital (O1, O2) areas. This setup allows an anatomic consistency across all 25 AD subjects, which allows valid cross-subject comparisons. The electrode layout is the base node set that is used for subsequent graph-theoretic and fractal analysis.

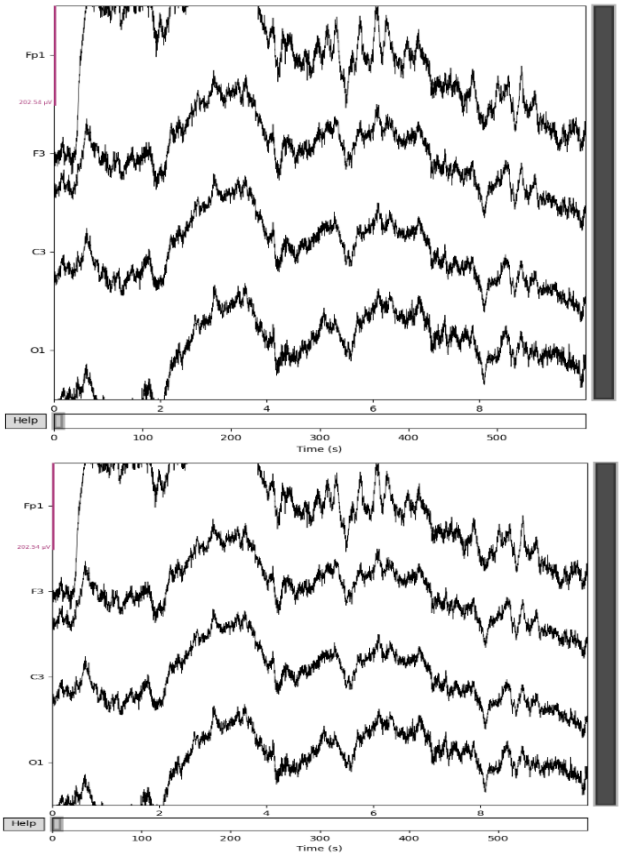


Fig. 2. A representative of EEG time series (10 seconds) from selected scalp electrodes of an Alzheimer's disease subject, illustrating multichannel electrical activity used for functional connectivity estimation.

Fig. 2 shows a 10-second resting state EEG recording from selected electrodes of the scalp in an Alzheimer's disease patient, showing simultaneously recorded multi-channel electrical activity. The traces correspond to different channels of the electrodes, showing the temporal fluctuation of brain electrical signals detected when the eyes are closed while resting. The signals show typical oscillatory, irregular patterns characteristic of neural activity in different regions of the brain. This is a multi-channel representation of raw input data from which functional connectivity matrices were estimated. The complexity and distinctiveness of the different channels indicate the heterogeneity of neural dynamics in AD.

III. Experimental Data Description

The study utilized resting-state EEG recordings from 25 Alzheimer's disease subjects obtained from the OpenNeuro repository (dataset ds004504) [23]. EEG signals were recorded at 500 Hz sampling frequency under eyes-closed resting-state conditions using 19 scalp electrodes positioned according to the international $10-20$ system. All recordings were stored in EEGLAB.set format and processed using an identical computational pipeline.

International $10-20$ electrode placement system (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz) is as follows. The same electrode configuration was used for all the subjects, thus ensuring consistency across analyses.

IV. Results and Discussion

This section provides the results of the proposed multi-scale fractal vulnerability analysis applied to EEG-derived Alzheimer's disease brain networks. The evolution of fractal structure and network connectivity as a function of network density is systematically explained, highlighting the identification of critical regimes where structural organization degrades. All results are reported for AD subjects only, with healthy control data used exclusively for qualitative reference in supplementary analyses. Consequently, the identified critical density should be interpreted as a characteristic property of the analyzed Alzheimer's disease cohort rather than a disease-specific biomarker.

Functional brain networks are constructed from EEG recordings by progressively thresholding the connectivity matrices across a predefined range of network densities. This density-based thresholding approach enables controlled sparsification of the network, allowing for a systematic analysis of the response of the network structure to a gradual decline in connections. As network density decreases, the total number of edges is reduced while preserving the strongest functional connections at the global level. At higher densities, networks remain fully connected and exhibit strong topological organization. However, as density decreases, networks begin to fragment, leading to the emergence of disconnected components. To avoid misleading effects arising from trivial disconnection, all subsequent analyses of fractal properties were performed exclusively on the largest connected component (LCC) of the network at each density level.

The size of the LCC decreases progressively with decreasing density. Notably, the reduction in LCC size is gradual rather than abrupt, indicating that topological disintegration occurs progressively rather than through a sudden percolation-like transition.

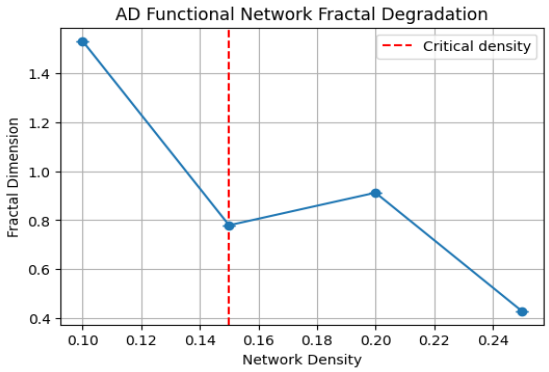


Fig. 3. Cohort-level mean fractal dimension as a function of network density.

Fig. 3 depicts the average fractal dimension of the cohorts as a function of network density $(0.10-0.25)$. The higher the density, the more stable the FD values are, which are in the range of $1.4-1.5$, indicating that the hierarchical organization is preserved. Around the critical density of 0.15, a strong drop is observed, suggestive of an abrupt change in the structure and not a gradual change. At lower densities, FD drops significantly to 0.4, indicating the almost full absence of fractal organization at

low density. The multi-scale analytical framework will be motivated by this non-monotonic, threshold-like behavior.

At higher densities, FD values remain relatively stable, representing preserved self-similar organization across multiple scales. As density decreases, FD begins to decline, reflecting a progressive loss of hierarchical structure and scale invariance.

Importantly, this decline is nonlinear. Rather than decreasing smoothly across all densities, FD exhibits a pronounced drop within a narrow density range. This behavior suggests the existence of a critical density regime where fractal organization becomes unstable.

The core finding of this study is the identification of a transitional density regime (approximately $15-25\%$) at which fractal structure collapses even though the network remains largely connected. This critical density is consistently observed across subjects and is marked by a sharp reduction in FD values. Importantly, this collapse occurs before severe topological fragmentation, as evidenced by the persistence of a substantial LCC at the same density levels. This decoupling between fractal degradation and network disconnection demonstrates that loss of fractal organization is not just merely a consequence of reduced network size but indicates a deeper structural vulnerability.

The critical density identified in this study represents a regime where the network transitions from a complex, multi-scale structure to a simpler, less organized topology. This phenomenon is not captured by conventional single-threshold analyses, underscoring the necessity of a multi-scale framework.

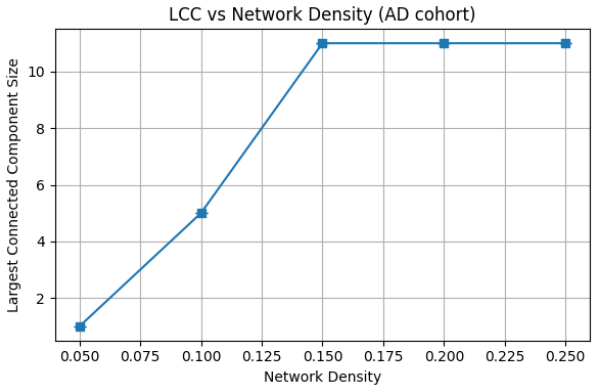


Fig. 4. Size of the largest connected component as a function of density.

Fig. 4 represents the size of the largest connected component plotted as a function of network density for the AD cohort, as shown in Fig. 4, and is found to increase in a monotonically increasing and smooth fashion from density $0.05-0.25$. Unlike the fractal dimension, the size of the LCC increases progressively with no sharp collapse or critical transition, ending at an LCC size of about $11-12$ nodes at a density of $0.15-0.175$. This smooth trajectory is in contrast to the steep FD drop at the same density range, which clearly indicates that connectivity and fractal organization deteriorate at a completely different rate. The finding that there is a difference between the FD and LCC trajectories lends credence to the hypothesis that fractal collapse is an earlier and more sensitive indicator of structural instability.

Unlike fractal dimensions, LCC size increased smoothly with density and did not exhibit abrupt collapse. Prominently, at the density where FD reached its minimum, the LCC remained large and stable across subjects. Progressive fragmentation is observed at lower densities, thus indicating increased network vulnerability.

To further explicate the relationship between structural complexity and connectivity, FD and LCC size were jointly analyzed across a range of densities. Even when both metrics decrease with decreasing density, the rates of decline differ substantially. LCC size decreases gradually across densities, whereas FD exhibits a steeper decline within a narrow range. This divergence indicates that fractal collapse precedes major connectivity loss, highlighting FD as a sensitive early marker of structural instability.

This result supports the hypothesis that fractal measures capture aspects of network organization that are not reflected by purely topological metrics. In the context of AD, this suggests that pathological changes may first manifest as disruptions in hierarchical organization rather than outright disconnection. To assess the robustness of the observed patterns, results were aggregated across all AD subjects by computing the mean and standard deviation of FD and LCC at each density level. The cohort-level trends are summarized in Table I and visualized in Fig. 5. The mean FD curve exhibits a clear inflection point corresponding to the critical density identified at the individual level. The associated standard deviation increases near this region, indicating heightened inter-subject variability during the transition phase. This increased variability suggests that the critical regime represents a structurally unstable state where individual networks diverge in their response to sparsification.

TABLE I. COHORT LEVEL MEAN ± STANDARD DEVIATION OF FRACTAL DIMENSION (FD) AND LARGEST CONNECTED COMPONENT (LCC) ACROSS NETWORK DENSITIES

	Threshold	FD (mean ± SD)	LCC (mean ± SD)
0	0.05	Nan	1.0
1	0.1	1.53	5.0
2	0.15	0.779	11.0
3	0.2	0.912	11.0
4	0.25	0.43	11.0

Table I shows the mean FD and LCC values across five density levels $(0.05-0.25)$ for the AD cohort, where the value of FD is not defined at the lowest density because the LCC size is too small. We observe a critical transition at density 0.15, where FD rapidly changes to 0.779, while LCC remains mostly constant at 11.0 nodes, thus signifying that fractal organization and connectivity break down in different ways.

Substantial FD reduction was observed within the intermediate density regime (approximately $15-25\%$), indicating structural instability across this sparsification range.

Traditional network analyses often rely on metrics computed at a single threshold or density, implicitly assuming stability across scales. To contextualize our findings, it is noted that FD values computed at a fixed density would fail to capture the observed non-monotonic behavior and critical transition. By

contrast, the proposed multi-scale approach reveals continuous evolution of structural complexity and identifies regimes of instability that are invisible to single-scale analyses. This highlights a fundamental limitation of fixed-threshold methods in the study of neurogenerative disorders.

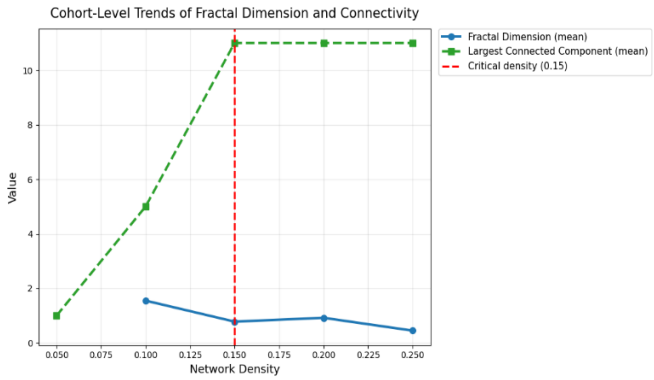


Fig. 5. Cohort-level trends.

Fig. 5 shows a dual-metric plot where the mean fractal dimension and the mean LCC size are plotted over the range of network densities, with the critical density of 0.15 indicated by a red dashed vertical line. The LCC curve increases sharply and levels off; the FD curve is relatively flat at high densities and drops quickly when approaching and reaching the critical density. The distance between the two curves is the largest exactly at the critical density, visually reinforcing that the two measures measure different but complementary aspects of network organization. The key evidence of the main claim of the study, that multi-scale fractal analysis reveals instability regimes that are not captured by single threshold connectivity measures, is the joint visualization.

The observed early collapse of fractal organization in AD networks has important implications. Fractal structure reflects the brain's ability to integrate information across multiple scales. Its degradation suggests a loss of hierarchical coordination that may underlie cognitive dysfunction in AD. By identifying a density regime where this collapse occurs independently of gross disconnection, the proposed framework provides a principled way to detect subtle structural vulnerabilities that precede large-scale network failure.

A. Identification of a Critical Density

The simultaneous observation of 1) a sharp reduction in fractal dimension and 2) preserved global connectivity identifies a critical density ($\sim 15-25\%$) at which structural self-similarity degrades before global network disconnection. This indicates that fractal organization is more sensitive to connectivity loss than conventional topological measures.

This study demonstrates that Alzheimer's disease EEG networks exhibit multi-scale structural vulnerability that cannot be captured by single-threshold analyses. The identification of a critical density at which fractal structure collapses while global connectivity remains intact suggests a distinct regime of pathological instability.

Importantly, the emergence of undefined fractal dimensions at very low densities is not a methodological artifact but a reflection of genuine network fragmentation. By explicitly restricting fractal analysis to the largest connected component, we ensure that observed changes reflect intrinsic network organization rather than trivial disconnection effects.

The proposed framework provides a systematic approach to threshold selection and highlights fractal dimension as a sensitive marker of early structural disintegration. This methodology is extendable to larger datasets and other neurodegenerative conditions.

The present study focuses exclusively on Alzheimer's disease networks and therefore does not establish whether the observed critical density is unique to Alzheimer's disease or reflects a broader property of EEG-derived functional networks. Future work incorporating healthy control cohorts will be necessary to determine disease specificity and evaluate the potential diagnostic relevance of the identified density regime.

V. LIMITATIONS

Several limitations should be acknowledged. First, the study utilized a relatively small Alzheimer's disease cohort obtained from a public EEG repository. Second, healthy control networks were not incorporated into the primary quantitative analysis, limiting conclusions regarding disease specificity. Third, functional connectivity was estimated using Pearson correlation, which captures linear relationships only. Future studies involving larger cohorts, multimodal neuroimaging data, and alternative connectivity measures may provide a more comprehensive characterization of multi-scale network vulnerability. Fourth, the present study employed minimally processed EEG recordings and did not incorporate dedicated artifact-removal procedures; therefore, the influence of residual artifacts cannot be completely excluded.

VI. CONCLUSION

In this study, we introduced a multi-scale fractal vulnerability framework to investigate the structural organization of EEG-derived functional brain networks in Alzheimer's disease from a graph-theoretic perspective. By systematically varying the density of the network and by integrating fractal dimension analysis with connectivity measures, we moved beyond conventional single-threshold approaches and demonstrated that Alzheimer's disease networks exhibit scale-dependent structural instability, which cannot be captured by traditional graph metrics. A core finding in this study is the transitional density regime ($\sim 15-25\%$), at which fractal organization collapses abruptly, even though global connectivity is preserved as measured by the largest connected component. There is a separate distinction between fractal degradation and topological disconnection. It reveals that loss of hierarchical, self-similar organization precedes network disintegration. Subtle and minute organizational breakdowns cannot be detected by pure topological metrics. In conclusion, this work demonstrates that multi-level fractal analysis reveals a distinct and early form of network vulnerability in Alzheimer's disease. By shifting from single-scale, static metrics to multi-scale, dynamic analysis, we have provided a new insight into the concept of structural vulnerability.

Reproducibility Statement

The entire analysis procedure, from EEG loading, functional connectivity estimation, density-controlled thresholding, graph construction, to the extraction of the largest connected component and calculation of fractal dimension, was applied equally among all the subjects. The parameters used and the density levels in the study are all explicitly explained to ensure that the proposed framework can be reproduced.

Data Availability Statement

All the EEG data taken in this study are from the publicly available OpenNeuro repository (dataset ds004504). All the data were fully anonymized before public release, and no personal information was accessible to authors. All the analyses were conducted in accordance with ethical guidelines and regulations involving human data.

Conflict of Interest

The authors declare that there are no commercial or monetary relationships that could be construed as a potential conflict of interest in the conduct of this research.

References

[1] Ponvel, P., Shahar, S., Singh, D.K.A. et al. The cost effectiveness of a multidomain intervention on physical, cognitive, vascular, dietary and psychosocial outcomes among community dwelling older adults with cognitive frailty in Malaysia: The AGELESS Trial. Alz Res Therapy 17, 101 (2025). https://doi.org/10.1186/s13195-025-01722-w

[2] T. T. Tin, W. J. Qian, A. Aitizaz, A. O. Salau, O. M. Ikumapayi, and S. A. Afolalu, "Towards the Hybrid Approach for Predicting Stroke Risk: A Feature Augmented Model," International Journal of Advanced Computer Science and Applications, vol. 16, no. 9, 2025, doi: https://doi.org/10.14569/ijacsa.2025.0160993.

[3] Maity, S., Alrubayan, M., Khan, M. M., & Pradhan, P. (2025). Alterations of brain tissue structural complexity and disorder in Alzheimer's disease (AD): Fractal, multifractal, fractal transformation, and disorder strength analyses (arXiv:2512.07061). arXiv https://arxiv.org/abs/2512.07061

[4] Yusuf, A. M., Yener Ilce, B., Alhaj, H. A., & Hamoudi, R. (2025). α-Synuclein in Parkinson's disease: A central point of convergence with depression. *npj Parkinson's Disease, 11*, Article 329. https://doi.org/10.1038/s41531-025-01167-w

[5] Thangaraj, C., Easwaramoorthy, D. Generalized fractal dimensions-based comparison analysis of edge detection methods in CT images for estimating the infection of COVID-19 disease. Eur. Phys. J. Spec. Top. 231, 3717–3739 (2022). https://doi.org/10.1140/epjs/s11734-022-00651-1

[6] Thangaraj, C., Easwaramoorthy, D., Muhiuddin, G., Selmi, B., & Kulish, V. (2025). Fractal based automatic detection of complexity in COVID-19 X-ray images. *Expert Systems*, *42*(1), e13497.

[7] Rong, Q., Thangaraj, C., Easwaramoorthy, D. et al. Multifractal based image processing for estimating the complexity of COVID-19 dynamics. Eur. Phys. J. Spec. Top. 230, 3947–3954 (2021). https://doi.org/10.1140/epjs/s11734-021-00336-1

[8] He, S., Thangaraj, C., Easwaramoorthy, D. et al. Multifractal analysis on age-based discrimination in X-ray images for sensing the severity of COVID-19 disease. Eur. Phys. J. Spec. Top. 231, 3663–3671 (2022). https://doi.org/10.1140/epjs/s11734-022-00615-5

[9] Jebali, H., Golmankhaneh, A.K. Fractal functions, self-similar measure and fractal dimensions on the Sierpiński gasket. Arab. J. Math. (2025). https://doi.org/10.1007/s40065-025-00575-3

[10] Pirici, D., Mogoanta, L., Ion, D. A., & Kumar-Singh, S. (2024). Fractal Analysis in Neurodegenerative Diseases. Advances in neurobiology, 36, 365–384. https://doi.org/10.1007/978-3-031-47606-8_18

[11] Díaz Beltrán, L., Madan, C. R., Finke, C., Krohn, S., Di Ieva, A., & Esteban, F. J. (2024). Fractal Dimension Analysis in Neurological Disorders: An Overview. Advances in neurobiology, 36, 313–328. https://doi.org/10.1007/978-3-031-47606-8_16

[12] Averna, A., Coelli, S., Ferrara, R., Cerutti, S., Priori, A., & Bianchi, A. M. (2023). Entropy and fractal analysis of brain-related neurophysiological signals in Alzheimer's and Parkinson's disease. *Journal of Neural Engineering, 20*(5). https://doi.org/10.1088/1741-2552/acf8fa

[13] Ziukelis, E. T., Mak, E., Dounavi, M. E., Su, L., & T O'Brien, J. (2022). Fractal dimension of the brain in neurodegenerative disease and dementia: A systematic review. Ageing research reviews, 79, 101651. https://doi.org/10.1016/j.arr.2022.101651

[14] Li, P., Yu, L., Lim, A. S. P., Buchman, A. S., Scheer, F. A. J. L., Shea, S. A., Schneider, J. A., Bennett, D. A., & Hu, K. (2018). Fractal regulation and incident Alzheimer's disease in elderly individuals. Alzheimer's & dementia : the journal of the Alzheimer's Association, 14(9), 1114–1125. https://doi.org/10.1016/j.jalz.2018.03.010

[15] Davidson, J. M., Zhang, L., Yue, G. H., & Di Ieva, A. (2024). Fractal dimension studies of the brain shape in aging and neurodegenerative diseases. In A. Di Ieva (Ed.), *The fractal geometry of the brain* (2nd ed., Vol. 36, pp. 329–363). *Advances in Neurobiology*. Springer. https://doi.org/10.1007/978-3-031-47606-8_17

[16] Dharia, S. Y., Liu, Q., Smith, S. D., & Valderrama, C. E. (2025). A novel approach for the early identification of genetic risk factors for Alzheimer's disease using EEG and psychometric data. *IEEE Journal of Biomedical and Health Informatics*. Advance online publication. https://doi.org/10.1109/JBHI.2025.3639217

[17] Liu, T. C., Zheng, T., Duan, R., Zhu, L., & Zhang, Q. G. (2020). On the Biomarkers of Alzheimer's Disease. Advances in experimental medicine and biology, 1232, 409–414. https://doi.org/10.1007/978-3-030-34461-0_52

[18] Chen, L., Zhou, X., Qiao, Y., Wang, Y., Zhou, Z., Jia, S., Sun, Y., & Peng, D. (2025). The impact of Alzheimer's disease on cortical complexity and its underlying biological mechanisms. Brain research bulletin, 225, 111320. https://doi.org/10.1016/j.brainresbull.2025.111320

[19] Honea, R. A., Hunt, S., Lepping, R. J., Vidoni, E. D., Morris, J. K., Watts, A., Michaelis, E., Burns, J. M., & Swerdlow, R. H. (2023). Alzheimer's disease cortical morphological phenotypes are associated with TOMM40'523-APOE haplotypes. Neurobiology of aging, 132, 131–144. https://doi.org/10.1016/j.neurobiolaging.2023.09.001

[20] Ashraf, G., McGuinness, M., Khan, M. A., Obtinalla, C., Hadoux, X., & van Wijngaarden, P. (2023). Retinal imaging biomarkers of Alzheimer's disease: A systematic review and meta-analysis of studies using brain amyloid beta status for case definition. Alzheimer's & dementia (Amsterdam, Netherlands), 15(2), e12421. https://doi.org/10.1002/dad2.12421

[21] Vicchietti, M. L., Ramos, F. M., Betting, L. E., & Campanharo, A. S. L. O. (2023). Computational methods of EEG signals analysis for Alzheimer's disease classification. Scientific reports, 13(1), 8184. https://doi.org/10.1038/s41598-023-32664-8

[22] Nicastro, N., Malpetti, M., Cope, T. E., Bevan-Jones, W. R., Mak, E., Passamonti, L., Rowe, J. B., & O'Brien, J. T. (2020). Cortical Complexity Analyses and Their Cognitive Correlate in Alzheimer's Disease and Frontotemporal Dementia. Journal of Alzheimer's disease: JAD, 76(1), 331–340. https://doi.org/10.3233/JAD-200246

[23] OpenNeuro. (n.d.). *EEG recordings from Alzheimer's disease, frontotemporal dementia and healthy subjects (Dataset ds004504)*. OpenNeuro. https://openneuro.org/datasets/ds004504

Progression-Aware Temporal Graph Transformer for Reliable Chronic Kidney Disease Trajectory Prediction

Dr. Roshan D. Suvaris[1], Dr. Padmavathy E.[2], Dilfuza Akabirkhodjaeva[3],
Dr. T. K. Rama Krishna Rao[4], Dr. R. Sindhu[5], Farrukh Sobia[6], Elangovan Muniyandy[7], Aseel Smerat[8]
Nitte (Deemed to be University)-NMAM Institute of Technology (NMAMIT), Department of MCA, Nitte, India, 574110[1]
Assistant Professor, Department: Mathematics, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India[2]
Department: System Analysis and Mathematical Modelling, The University of World Economy and Diplomacy, Uzbekistan[3]
Professor, Department of Computer Science and Engineering, Koneru Lakshmaih Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[4]
Assistant Professor, Department of Mathematics, Dr. N.G.P Arts and Science College, Coimbatore-641048, Tamil Nadu, India[5]
Department of Health Informatics, College of Public Health and Tropical Medicine-Jazan University, Jazan, Kingdom of Saudi Arabia[6]
Department of Biosciences, Saveetha School of Engineering-Saveetha Institute of Medical and Technical Sciences, Chennai, India - 602 105[7]
Faculty of Educational Sciences, Al-Ahliyya Amman University, Amman, 19328, Jordan[8]

Abstract*—Chronic Kidney Disease (CKD) is a progressive and irreversible condition that requires early prediction of renal deterioration for effective clinical intervention. Existing studies based on static machine learning and conventional deep learning models fail to capture temporal dependencies, evolving biomarker interactions, and longitudinal disease progression patterns, leading to limited predictive reliability in real-world clinical settings. To address these limitations, this study proposes a Progression-Aware Temporal Graph Transformer for reliable CKD trajectory prediction using the Chronic Renal Insufficiency Cohort (CRIC) dataset comprising 5,625 patients with longitudinal clinical follow-ups and repeated biomarker measurements including eGFR, creatinine, albumin, blood pressure, glucose, and HbA1c. The proposed framework integrates a temporal transformer encoder for sequential patient representation, a dynamic biomarker graph to model evolving renal interactions, and a multi-scale temporal attention mechanism to capture both short- and long-term deterioration patterns. The model is implemented using Python with PyTorch and executed on GPU-based computational infrastructure for efficient training and inference. Experimental results demonstrate that the proposed model achieves 95.0% accuracy, 94.2% precision, 95.1% recall, and 94.6% F1-score, improving performance by approximately 2–4% over state-of-the-art baselines such as Random Forest, KNN, GNN, and 1D CNN models. Additionally, survival analysis yields a Concordance Index of 0.952, confirming strong risk-ranking capability for dialysis onset prediction. The framework also maintains robust performance under noisy and missing data conditions, demonstrating strong generalization ability. In conclusion, the proposed model provides an interpretable, scalable, and clinically reliable solution for CKD progression forecasting, enabling early intervention, personalized treatment planning, and improved renal outcome prediction in clinical decision-support systems.

Keywords*—Chronic Kidney Disease; temporal graph learning; disease progression prediction; longitudinal healthcare modeling; trajectory prediction

I. Introduction

Chronic Kidney Disease is a chronic, progressive and life threatening condition where the kidneys function over time [1]. It impacts millions of people worldwide and is a significant public health problem because of its close link with cardiovascular disease, ESRD, and higher death rate [2] [3]. Many cases of early stage CKD are not diagnosed and are not treated because they do not have any symptoms [4]. Therefore, it is important to monitor renal biomarkers like estimated glomerular filtration rate (eGFR), serum creatinine, blood pressure, glucose level and albuminuria continuously to detect progression of disease before irreversible kidney damage [5]. Early prediction of CKD progression is an important tool for improving patient outcomes by enabling timely clinical interventions, patient-tailored treatment and avoiding dialysis dependency [6].

Although significant progress has been made in the prediction of CKD using machine learning and deep learning, there remain several key limitations in the current approaches [7]. Most of the conventional models are based on the binary classification of the presence or absence of CKD, and do not account for the temporal patterns of renal deterioration. Such techniques typically analyze each clinical variable separately and fail to account for the changing relationships among the biomarkers over time for individual patients [8]. Moreover, many of the current methods are not capable of temporal reasoning, which limits their usefulness in modeling a progressive decline in kidney function over the long term [9]. Existing models also lack the flexibility to adapt to different

disease states as the disease progresses, due to the use of fixed graph structures and shallow attention mechanisms [10]. Furthermore, the lack of interpretation and lack of clinical transparency decreases the trust of physicians and hampers the direct use in clinical practice.

Motivated by these challenges, this study proposes a novel framework entitled Progression-Aware Temporal Graph Transformer for Reliable Chronic Kidney Disease Trajectory Prediction. The proposed model utilises patient trajectories over time, along with the interaction of biomarkers that is dynamic, to predict future patterns of renal deterioration. The temporal transformer encoding is used to model the sequential clinical evolution, and the progression-aware graph learning models the adaptive interdependencies between biomarkers at various clinical stages. Furthermore, multi-scale temporal attention is used to boost long-term progression reasoning, and make the clinical interpretation more interpretable. The proposed framework is able to predict CKD trajectories instead of the static classification, which enables early intervention, prevents dialysis, and assists with nephrology decisions for clinical prognosis.

A. Problem Statement

Existing prediction approaches for CKD are mostly static prediction of the disease stage and are not well suited to the longitudinal modelling of renal deterioration in the repeated visit of patients [11]. Traditional machine learning and deep learning methods focus on the clinical features themselves rather than dynamically evolving relationships between the biomarkers, like the interactions between creatinine, glucose, blood pressure and eGFR over time [12]. In addition, most graph-based methods use fixed graph structures, which are unable to evolve to different phases of CKD. The other major drawback is the lack of current models which can predict future renal decline, dialysis risk or disease course in advance. These constraints limit the clinical validity and early intervention potential, and underscore the importance of progression-aware temporal learning frameworks that can model evolve temporal biomarkers and long-term CKD trajectory forecasting.

B. Key Contribution

- A progression-aware temporal graph transformer framework is proposed to model longitudinal clinical trajectories for reliable Chronic Kidney Disease progression prediction.
- Dynamic biomarker graph learning is introduced to capture evolving interdependencies among renal biomarkers across different stages of disease progression.
- A multi-scale temporal attention mechanism is developed to simultaneously learn short-term fluctuations and long-term renal deterioration patterns from sequential patient records.
- The proposed framework performs future trajectory forecasting by predicting estimated glomerular filtration rate decline and Chronic Kidney Disease stage transitions using longitudinal CRIC patient data.
- Extensive experimental evaluation and comparative analysis demonstrate improved predictive reliability, temporal reasoning capability, and clinical interpretability over existing machine learning and deep learning approaches for CKD prognosis.

The remainder of this study will be structured as follows: Section II will contain a summary of the study that is currently available. Section III is about the model proposed and its implementation. The Section IV entails experiments and discussion that accompanied them. Last conclusion and future work in Section V.

II. Related Works

Lei et al. [13] performed a systematic review and meta-analysis to investigate the efficacy of ML techniques in predicting the development of kidney disease. This research involved the examination of 15 articles, where 6 studies, including 12,534 patients, were considered in the meta-analysis. According to the results, ML models had sufficient predictive performance, with an AUC of the HSROC of 0.87. The ML techniques exhibited high specificity and average sensitivity concerning the identification of disease progression. Moreover, the subgroup analysis for CKD and IgA nephropathy demonstrated a decent predictive accuracy as well.

Islam et al., [14] Examined the use of machine learning and predictive modeling in early detection of CKD. In this research, 25 different clinical variables were analyzed by means of feature selection to determine the most important predictors. This helped in reducing the data set to about 30 percent of the original features. Twelve different supervised machine learning classifiers were considered for predicting CKD. It was found that the XGBoost classifier gave the best results with accuracy, precision, recall, and F1 score of 98.3, 98, 98, and 98, respectively.

Nagawa et al., [15] proposed a new method of CKD diagnosis through ultra-wide-field fundus examination as opposed to conventional laboratory examination. They used convolutional neural networks to identify discriminative retinal features that are related to kidney disease progression. The research used a huge imaging dataset and showed a high sensitivity and specificity, which reveals a good diagnostic ability. One of the strengths is that its screening procedure is non-invasive such that CKD can be detected during eye check-ups. However, the approach depends on specialized imaging equipment and lacks integration with structured clinical data, which may restrict its scalability and applicability in resource-constrained healthcare environments.

Tawsik Jawad et al.,[16] investigated the use of XAI with the ensemble method for predictive analysis of CKD. In this study, a predictive model was created based on accurate predictions by utilizing clinical biomarkers to increase transparency in the decision-making process through XAI methods. Using the suggested framework, the most relevant features were determined, which impact the prediction of CKD, as well as the reasons for each prediction, making it possible for doctors to make informed decisions regarding treatment. The results of the investigation showed that integration of ensemble learning and XAI improves prediction accuracy and early detection of CKD.

Arif et al., [17] aimed to enhance CKD prediction with an enhanced preprocessing and feature selection pipeline.. They used the Boruta algorithm, feature relevance analysis, and anomaly detection to identify predictors that are stable and significant. The method was experimented on clinical data and found to be highly classified, which indicates the significance of preprocessing in a medical machine learning system. Nonetheless, the research has a limitation of a small sample size and lacks a wide range of population. The lack of external validation further limits its use in actual healthcare context where heterogeneity of the data is a crucial criterion.

Al-Momani et al., [18] carried out a comparative analysis to compare various machine learning algorithms in the prediction of CKD. The research analyzed Artificial Neural Networks, Support Vector Machines, and K-Nearest Neighbors using clinical data. Of these, Artificial Neural Networks had the best accuracy, and therefore they have the potential to model nonlinear relations within CKD data. The research offers good information about the performance of the algorithms, but has shortcomings in terms of limited features and absence of long-term patient outcomes. These weaknesses limit the model in terms of capturing disease progression, and impact its strength in a real-world clinical usage.

Huamg et al., [19] introduced a predictable and computationally cost effective CKD forecasting framework with an optimized XGBoost model. The approach included Bayesian optimization using Optuna and explained using SHAP. The model was tested on structured electronic health records and performed better at predicting that at a lower computational cost. One of the benefits is accuracy/interpretability balance. But the framework is heavily dependent on quality structured data, which restricts its applicability to heterogeneous or unstructured clinical settings. It also does not incorporate relational feature learning to learn biomarker interactions.

The main limitation of the literature on CKD prediction is that researchers focus on fragmented modeling approaches, considering preprocessing, feature selection, prediction and explainability as individual elements, which results in poor generalization and weak performance in conditions of noisy or imbalanced clinical data. A lot of methods are based on manual preprocessing or single-level attention and explainability is commonly post-hoc and not connected to the learning process. Scalability and other clinical feasibility limitations are also added with imaging-based and hybrid models. The proposed Progression-Aware Temporal Graph Transformer predicts future Chronic Kidney Disease progression by learning longitudinal biomarker trajectories, dynamic temporal relationships, and evolving renal deterioration patterns, enabling accurate forecasting of eGFR decline, CKD stage transition, dialysis risk, and hospitalization probability.

Table I provides a comparative study of the current CKD prediction models, their methodologies, datasets, strengths, and limitations. Although most methods are very accurate and have better interpretability, most of them have shortcomings, including inability to scale to large data sets, reliance on structured data, no relational modeling, and lack of validation in a variety of clinical settings.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING CKD PREDICTION METHODS

Author	Methodology	Advantage	Disadvantage
(Ghosh & Khandoker,[20]	GBoost with SHAP, LIME	Accurate and interpretable predictions	Small imbalanced dataset limitations
Iftikhar et al.,[21]	Multiple ML model comparison	Comprehensive model evaluation.	No temporal progression modeling.
Dritsas & Trigka, 2022b [22]	Class balancing with ML	Extremely high prediction accuracy	No temporal progression modeling
Jawad et al. [23]	RF-based ML model	High accuracy and explainability	Limited feature interaction modeling
Arif et al. [17]	Boruta + preprocessing pipeline	Stable feature selection	Small dataset, no validation
Al-Momani et al. [18]	ANN, SVM, KNN comparison	Strong ANN performance	Limited features, no temporal data
Huang et al. [19]	XGBoost + SHAP + Optuna	Efficient and interpretable	Depends on structured data

III. PROPOSED PROGRESSION-AWARE TEMPORAL GRAPH LEARNING FOR EARLY CKD PROGRESSION FORECASTING

The proposed methodology proposes a Progression-Aware Temporal Graph Transformer framework for forecasting the progression of Chronic Kidney Disease in the CRIC dataset. The proposed framework is distinct from traditional CKD prediction models that make a single binary classification, as it captures the renal deterioration trajectory over multiple clinical visits. Longitudinal patient records with renal biomarkers, demographics and clinical measures are preprocessed initially with missing value imputation, temporal alignment, feature normalization and progression-aware balancing. Sequential patient trajectories are then built to ensure that the disease evolution patterns are maintained visit wise. The proposed framework aims to learn the temporal biomarker dependencies and the behavior of the changes over time from repeated patient visits using a Temporal Transformer Encoder. Biomarkers are shown as dynamic graph nodes to reflect the changing nature of the physiological interactions, and clinical relationships as adaptive temporal edges. The main novelty of the study lies in this progression-aware graph construction because the interactions between biomarkers are dynamic and change throughout the different clinical stages of renal deterioration and in function of the severity of the CKD. For instance, glucose and HbA1c have a greater effect in early diabetic nephropathy, while creatinine, albumin, and eGFR dominate during the advanced stages of CKD progression. Moreover, a Multi-Scale Temporal Memory Attention mechanism is introduced to learn simultaneously both short-term fluctuations in biomarkers and long-term patterns of renal deterioration acceleration. Finally, a forecasting head forecasts future eGFR decline, CKD stage transition, dialysis onset probability and renal hospitalization risk over a number of time horizons. The proposed methodology addresses the temporal reasoning, progression consistency, adaptive biomarker learning and clinical interpretability challenges, which are not addressed by current static CKD prediction methods. Fig. 1 shows overall workflow architecture of proposed model.

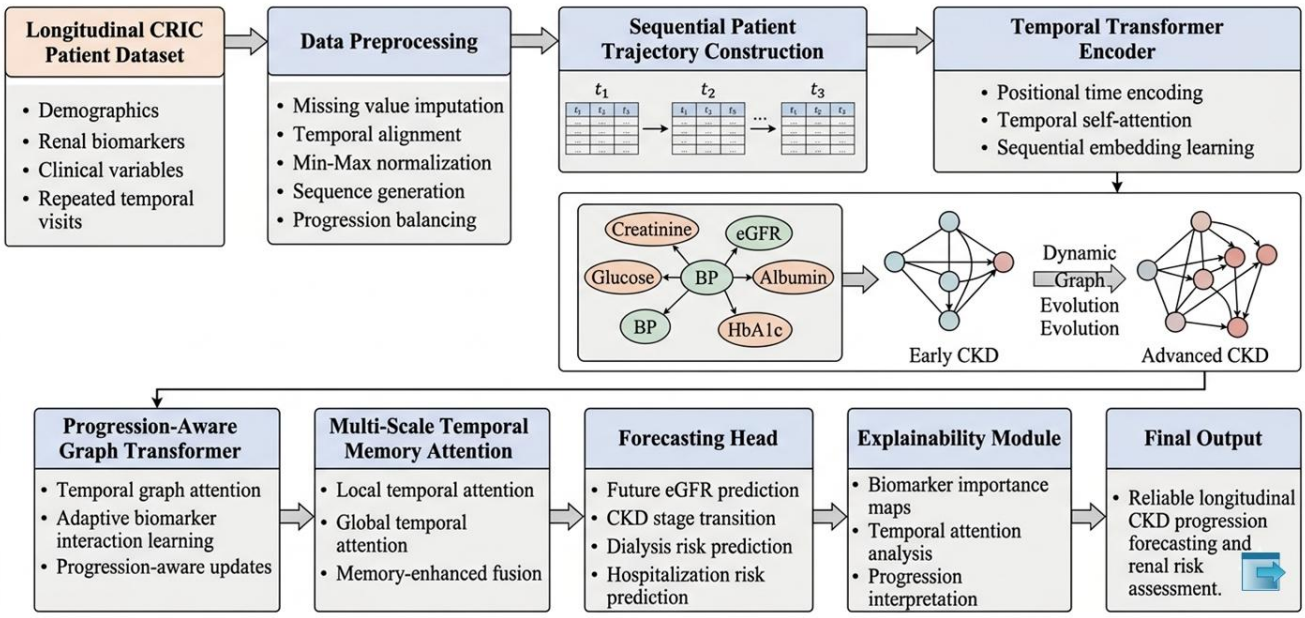


Fig. 1. Proposed progression-aware temporal graph learning for early ckd progression forecasting framework.

A. Data Collection

The dataset analyzed in this study is the Chronic Renal Insufficiency Cohort Study (NIDDK CRC) longitudinal clinical data set from the NIDDK Central Repository [24]. The CRIC data set includes clinical data from some 5,625 patients with Chronic Kidney Disease, who have been followed longitudinally over time since 2003. The data comprises annual clinical visits as well as evaluations at 6-month intervals, offering detailed time trajectories for individual patients for renal progression analysis. There are approximately 400+ clinical variables and laboratory parameters available, such as demographic, renal biomarkers, cardiovascular parameters, medication history, lifestyle and comorbidity data.

Major demographics like age, gender, ethnicity and body mass index (BMI) are chosen for this study. Serum creatinine, estimated glomerular filtration rate (eGFR), and urinary albumin-to-creatinine ratio (UACR), along with albumin, hemoglobin and blood urea nitrogen, are all important renal biomarkers that are used to monitor kidney function deterioration. Other clinical parameters like blood glucose, SBP, DBP, HbA1C, diabetes status and cardiovascular parameters are also included. The proposed framework is predicted to cover three important clinical outcomes: future eGFR decline, transition to CKD stage and estimation of dialysis risk. The CRIC dataset is ideal for modelling the temporal progression of the disease and predicting CKD trajectories due to repeated patient visits and longitudinal measurements of biomarkers.

B. Data Pre-Processing

Mode imputation was used to fill in categorical features, whereas missing values for numerical biomarkers like creatinine, hemoglobin, albumin, blood pressure, and eGFR were imputed through median imputation, which involves calculating the median of the available values of each feature. In cases where a single value is missing within a longitudinal patient's trajectory, temporal interpolation is utilized to maintain temporal consistency between neighboring visits. This is stated in Eq. (1) below.

$$X_t = X_{t-1} + \frac{t-t_{-1}}{t_{+1}-t_{-1}}(X_{t+1}-X_{t-1}) \quad (1)$$

where, X_t represents interpolated biomarker value at time t, X_{t-1} and X_{t+1} represent nearest available observations before and after the missing visit, respectively, and t_{-1} and t_{+1} denote their corresponding timestamps. This formulation performs linear temporal interpolation to estimate missing biomarker measurements while preserving the temporal continuity of longitudinal patient trajectories.

Since patient visits occur at irregular time intervals, all visits are temporally aligned and converted into ordered longitudinal sequences based on visit timestamps. This allows disease progression analysis in terms of patient-specific temporal trajectories to be constructed. To stabilize transformer training and avoid feature dominance, Min-Max normalization is applied to all numerical biomarkers as (2).

$$Z_{norm} = \frac{Z-Z_{min}}{Z_{max}-Z_{min}} \quad (2)$$

where, Z is the feature value, Z_{min} is the minimum, and Z_{max} is the highest of that feature. Moreover, class imbalance due to the scarcity of severe CKD progression events is corrected by introducing Temporal SMOTE, focal loss optimization and progression-aware weighting methods for better prediction of rare events in CKD progression and robust temporal generalization.

Approximately 5,000 patient trajectories were retained following the preprocessing, temporal sequence generation and the elimination of longitudinal data of patients with incomplete data. Nearly balanced longitudinal training samples were created by applying Temporal SMOTE and progression-aware resampling to overcome the progression-event imbalance in order to ensure reliable CKD trajectory forecasting.

C. Sequential Patient Trajectory Construction

The patient trajectory construction problem in the longitudinal CRIC dataset is solved by the sequential patient trajectory construction method, which will be a structured temporal evolution sequence of diseases of patients in the CRIC that can be used for progression-aware learning. Unlike traditional CKD prediction approaches that rely on isolated clinical snapshots, this framework models continuous renal deterioration across repeated follow-up visits, enabling a more realistic representation of disease progression dynamics. Each patient P_i is represented as an ordered trajectory in (3).

$$P_i = \{Xi^1, X_i^2, Xi^3, \dots, X_i^T\} \quad (3)$$

where, X_i^T denotes the biomarker feature vector recorded at visit t and T represents the total number of available longitudinal assessments.

Each temporal biomarker vector is defined as $Xi^{\mathrm{t}} = [Cr_{\mathrm{t}}, eGFR_{\mathrm{t}}, UACR_{\mathrm{t}}, Glucose_{\mathrm{t}}, HbA1c_{\mathrm{t}}, BP_{\mathrm{t}}, Albumin_{\mathrm{t}}, BMI_{\mathrm{t}}]$, capturing key renal, metabolic, and physiological indicators. Here, $\mathrm{Cr_t}$ represents serum creatinine, $\mathrm{eGFR_t}$ denotes estimated glomerular filtration rate, $\mathrm{UACR_t}$ indicates urine albumin-to-creatinine ratio, $\mathrm{Glucose_t}$ and $\mathrm{HbA1c_t}$ reflect glycemic status, $\mathrm{BP_t}$ captures blood pressure, $\mathrm{Albumin_t}$ represents serum albumin levels, and $\mathrm{BMI_t}$ denotes body mass index.

This structured representation retains the longitudinal dependencies and allows learning of the patterns of progressive deterioration of biomarkers associated with CKD progression. For example, rise in creatinine along with a fall in eGFR would mean a deterioration in renal function whereas rise in glucose and HbA1c would mean progression of diabetic nephropathy.

To further capture temporal dynamics, transition variations between consecutive visits are computed as $\Delta Xi^{\mathrm{t}} = Xi^{\mathrm{t+1}} - Xi^{\mathrm{t}}$, representing biomarker change over time. This formulation allows to estimate the velocity of progression, deterioration rate and trends in temporal evolution. These sequential representations are then passed to the Temporal Transformer Encoder and Dynamic Temporal Graph Learning module for the purposes of CKD stage transition prediction, future eGFR forecasting and dialysis risk assessment.

D. Temporal Biomarker Embedding

The Temporal Biomarker Embedding stage aims to model the longitudinal renal progression behavior from a series of patient trajectories in CRIC. The proposed framework utilizes a Temporal Transformer Encoder to capture temporal dependencies, visit-wise disease progression and evolving biomarker interactions across various follow-up visits, unlike the conventional CKD prediction models, which process independent clinical records, where there is no temporal continuity. This stage will allow the model to learn the pattern of the renal biomarkers that occur over time and the nature of their contribution to the subsequent worsening of CKD. First temporal biomarker vectors are merged with positional time encoding to retain the temporal order of visits to the patient. The temporal input representation is formulated as (4).

$$Z_t = X_t + PE_t \quad (4)$$

where, Z_t denotes the temporal input embedding at visit t, X_t represents the biomarker feature vector at time t, and PE_t denotes positional temporal encoding used to preserve visit order information. The positional encoding helps the transformer to distinguish the different observation times of the kidney, maintaining the order of the progression of renal events across the patient's history.

The Temporal Transformer Encoder is then used to calculate the contextual temporal representations based on self-attention learning. The framework is able to capture the temporal self-attention, which means it knows how past renal states affect future kidney deterioration. For instance, if a patient's creatinine and UACR continue to rise over multiple appointments and his or her eGFR continues to fall over time, the weights for these measures are larger at each visit, allowing for the accurate prediction of progression. The temporal embeddings generated capture short-term fluctuations in the biomarkers as well as long term patterns of renal decline. These learned representations are then passed to the Dynamic Temporal Graph Learning module for modelling the evolving biomarker relationships and progression-aware CKD forecasting.

E. Dynamic Temporal Biomarker Graph Construction

Progression-Aware Temporal Graph Transformer's main contribution is the Dynamic Temporal Biomarker Graph Construction module. The proposed approach dynamically models the evolving renal interactions over the course of the disease with longitudinal clinical visits, whereas the conventional CKD prediction models assume fixed relationships between the biomarkers throughout the disease. The nodes in each time step are the individual biomarkers, and the adaptive graph edges are a model of physiological and pathological relationships between biomarkers. This allows the framework to reflect the dynamics of renal interactions that are different across disease stages in the course of CKD. At each timestamp t, the temporal biomarker graph is represented as: $R_t = (O_t, E_t)$ where R_t denotes the dynamic biomarker graph at temporal visit t, O_t represents the set of biomarker nodes, and E_t denotes adaptive temporal edges representing biomarker interactions.

The renal and clinical variables included in the biomarker node set are creatinine, eGFR, glucose, blood pressure, albumin, HbA1c and UACR. Edges in the graph are the emerging physiological relationships between the biomarkers. For instance, glucose and HbA1c have more interactions in the early stage of diabetic nephropathy, and creatinine and albumin have more interactions in the later stage of renal deterioration. To model evolving biomarker relationships, temporal graph attention is computed as (5).

$$h_i^{(t+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(t)} W_t h_j^{(t)}\right) \quad (5)$$

where, $h_i^{(t+1)}$ denotes the updated representation of biomarker node i at the next temporal state, $N(i)$ represents neighboring biomarkers connected to node i, $\alpha_{ij}^{(t)}$ denotes temporal attention weight between biomarkers i and j, W_t represents learnable temporal transformation weights, $h_j^{(t)}$ denotes neighboring biomarker representations, and σ represents nonlinear activation. The adaptive graph evolution

allows the framework to learn the progression-aware renal dependency patterns between the different stages of CKD. As a result, the model reflects clinically relevant deterioration processes, changing interactions between diseases, and temporal renal progression patterns better than static graph architectures.

F. Progression-Aware Graph Transformer and Temporal Attention Mechanism

Proposed framework for the longitudinal prediction of CKD progression is based on the Progression-Aware Graph Transformer, which is the core learning architecture. The aim of this module is to learn dynamically the evolving relationship between biomarkers and detect clinically important patterns of renal deterioration along patient's temporal trajectories. The proposed transformer dynamically updates the graph attention to reflect the severity and progression rate of CKD and the temporal evolution of the disease, whereas the conventional graph neural networks make the static assumption of biomarker relations. This allows focusing on increasing biomarker levels and accelerating renal decline patterns that are very likely to impact future kidney decline.

The temporal attention coefficient is a dynamic measure of clinical relevance of the interactions between the biomarkers throughout the disease progression. Thus, during the early stages of CKD, glucose, HbA1c and BP are given more priority because of the renal effects of diabetes and hypertension. In advanced stages of CKD, greater emphasis is placed on UACR progression, deterioration of albumin, eGFR decline, and creatinine, as these markers reflect severe renal impairment.

The proposed attention mechanism also captures the progression velocity and temporal deterioration acceleration by highlighting the fast deterioration of biomarker trajectories at the different visits. This learning is progression aware, which allows the framework to identify patients who are improving from those who are getting worse quickly. Thus, the resulting graph representations maintain clinically meaningful patterns of temporal progression, improve the ability to predict long-term graph trajectories, and increase the accuracy of the prediction of future eGFR decline, progression to CKD stage and the risk of dialysis. Fig. 2 shows Progression-Aware Graph Transformer architecture.

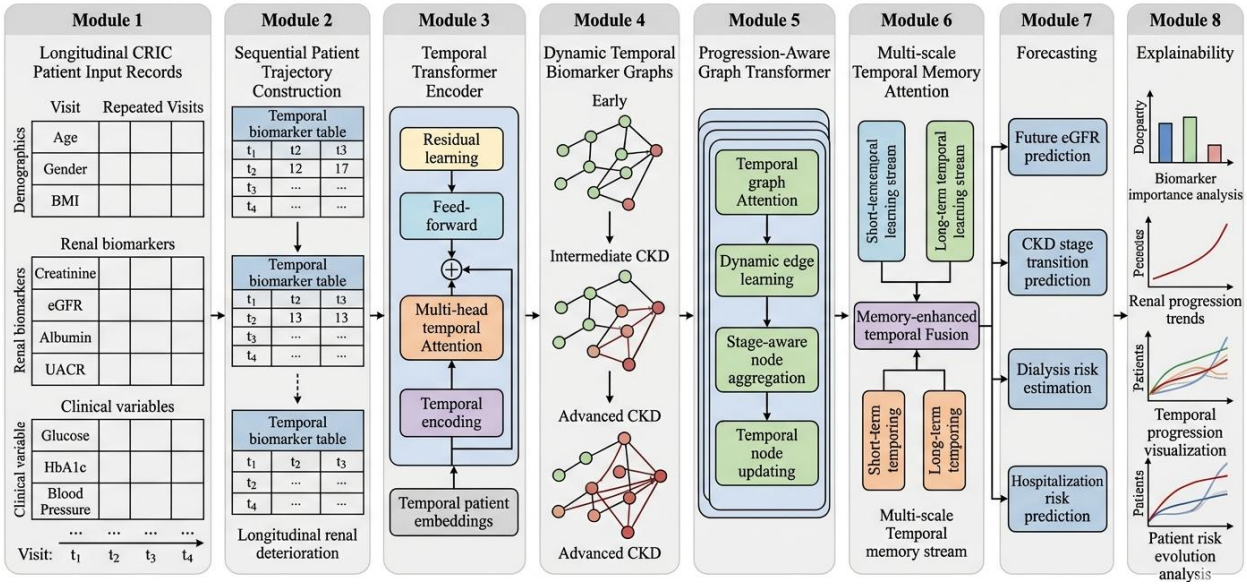


Fig. 2. Progression-aware graph transformer architecture.

G. Multi-Scale Temporal Attention and Memory Learning

The Multi-Scale Temporal Attention module aims to detect short-term biomarker fluctuations and identify long-term renal deterioration patterns from longitudinal trajectories of patients in the CRIC dataset. This module is a novel aspect of the proposed framework as CKD progression is a process that occurs over a long period and is accompanied by sudden deterioration events. In traditional temporal and sequential models, the learning of the local variations and the global evolution trends of the diseases are not performed simultaneously. To overcome this shortcoming, the proposed model combines local temporal attention, global temporal attention and memory-enhanced transformer blocks to enhance the progression-consistent CKD forecasting.

The local temporal attention first learns short-term fluctuations of the biomarkers between neighboring visits. This allows the model to detect abrupt renal deterioration signals like rapid decrease in eGFR, rise of creatinine, albuminuria, and abnormal changes in blood pressure. Simultaneously, the global temporal attention mechanism is able to learn long-term disease development patterns from the entire disease history, allowing for robust modelling of chronic renal decline patterns. The multi-scale temporal attention mechanism is formulated as (6).

$$M_t = \lambda L_t + (1 - \lambda) G_t \tag{6}$$

Where M_t denotes the final multi-scale temporal memory representation, L_t represents local temporal attention features capturing short-term variations, G_t denotes global temporal attention features representing long-term progression behavior, and λ represents adaptive weighting used to balance local and global temporal information. The memory-enhanced

transformer block then updates temporal disease representations using (7).

$$H_t' = Transformer(M_t, H_{t-1}) \quad (7)$$

where, H_t' denotes updated temporal memory representations, M_t represents multi-scale temporal features, and H_{t-1} denotes previous temporal memory states.

This architecture allows the framework to keep a continuity of progression, acquire the acceleration of deterioration patterns, and keep stability in forecasts for long prediction horizons. Thus, the model is able to achieve better consistency in the prediction of future eGFR, estimation of the CKD stage transitions and forecasting of dialysis risk.

H. Renal Progression Forecasting Head

The Renal Progression Forecasting Head is designed for multi-task longitudinal prediction of future renal outcomes based on the proposed framework, which infers temporal graph representations from the learned representations. The proposed forecasting head predicts future kidney deterioration trajectories across multiple clinical endpoints, as opposed to conventional CKD models, which only perform binary disease classification. This allows for early intervention planning, dialysis prevention and individualized renal risk assessment.

Multi-Scale Temporal Attention module provides progression-aware temporal embeddings to the forecasting architecture, which is then fed into fully connected prediction layers. The temporal representations are firstly aggregated with global temporal pooling, then followed by dense transformation layers for task specific renal outcome prediction. The output branches are separated for the future estimation of eGFR, prediction of CKD stages, risk of dialysis onset, and predicting renal hospitalization. The generalized forecasting function is formulated as (8).

$$\hat{Y}_{t+\Delta} = f(H_t') \quad (8)$$

where, $\hat{Y}_{t+\Delta}$ denotes predicted future renal outcomes after time interval Δ, H_t' represents progression-aware temporal patient embeddings, and $f(\cdot)$ denotes the forecasting network composed of dense prediction layers. Future eGFR prediction is represented as (9).

$$eGFR_{t+\Delta} = W_f H_t' + b_f \quad (9)$$

where, $eGFR_{t+\Delta}$ denotes predicted future eGFR value, W_f represents learnable forecasting weights, and b_f denotes bias parameters.

The forecasting head can forecast CKD stage transitions, the onset probability of dialysis and risk of renal hospitalization at 3-month, 6-month and 12-month horizons. This multi-task learning strategy improves progression consistency, enhances clinical interpretability, and enables comprehensive longitudinal renal deterioration forecasting.

I. Optimization Strategy, Explainability Module, and Computational Complexity Analysis

The proposed framework utilizes a Stability-Aware Optimization strategy to enhance the robustness of the forecasts, temporal consistency and stability in dynamic graph learning for modeling longitudinal CKD. For stable convergence and less overfitting, the network parameters are optimized by using the AdamW optimizer and cosine-learning rate scheduling for large temporal patient trajectories. To compensate for progression imbalance and rare events of renal deterioration, focal temporal loss is included to highlight hard progression samples and more rapid deterioration in CKD. The joint optimization objective is defined as (10).

$$L = \lambda_1 L_{forecast} + \lambda_2 L_{temporal} + \lambda_3 L_{graph} \quad (10)$$

where, $L_{forecast}$ represents renal outcome forecasting loss, $L_{temporal}$ denotes temporal consistency loss, L_{graph} represents dynamic graph relational learning loss, and λ_1, λ_2, and λ_3 represent adaptive optimization weights that control the relative importance of the three losses in the training process. The values of the latter three variables have been selected by empirically tuning them according to the model's validation performance in order to find an optimal compromise between forecasting accuracy, temporal consistency, and graph-learning stability. Sensitivity analysis of the weighting parameters is out of the scope of the present research.

To create biomarker importance maps, temporal attention visualizations, renal trajectory explanations and progression-risk interpretations, an explainability module is integrated. This enhances the clinical interpretability by focusing on biomarkers that are influential, like creatinine, eGFR decline, albuminuria and blood pressure deterioration, throughout the stage of disease progression.

The computational complexity analysis discusses the complexity of attention in transformer, scalability of dynamic graph attention, and longitudinal temporal processing. The proposed architecture is scalable in terms of temporal learning and is able to process complex patient trajectories, which may be large in CRIC, and changing graphs of biomarkers.

Algorithm 1: Progression-Aware Temporal Graph Transformer for CKD Progression Forecasting

Input:
- Longitudinal CRIC dataset D
- Patient trajectories P
- Biomarker features F
- Temporal visits T

Output:
- Future eGFR prediction
- CKD stage transition
- Dialysis risk estimation
- Hospitalization risk prediction

Begin

1. Load CRIC longitudinal patient dataset D
2. Data Preprocessing
   - For each patient record in D:
     - Handle missing numerical values using median imputation
     - Handle categorical attributes using mode imputation
     - Perform temporal interpolation for sequential visits
     - Normalize biomarker features using Min-Max normalization
   - End For

3. Temporal Sequence Construction
   For each patient Pi:
      Arrange visits chronologically
      Construct temporal trajectory:
         Pi = {X1, X2, X3, ..., Xt}
   End For
4. Temporal Biomarker Embedding
   For each temporal visit Xt:
      Add positional temporal encoding
      Generate temporal embeddings using Transformer Encoder
   End For
5. Dynamic Temporal Graph Construction
   For each timestamp t:
      Create graph Gt = (Vt , Et)
      Represent biomarkers as graph nodes
      Construct adaptive temporal edges between biomarkers
   End For
6. Progression-Aware Graph Attention
   For each biomarker node i:
      Compute temporal attention weights αij(t)
      Update node representations:
         hi(t+1) = σ(Σ αij(t) Wt hj(t))
   End For
7. Multi-Scale Temporal Memory Learning
   Learn local temporal attention features
   Learn global temporal attention features
   Fuse short-term and long-term representations
   Generate progression-aware memory embeddings
8. Forecasting Head
   Predict future eGFR values
   Predict CKD stage transitions
   Predict dialysis onset probability
   Predict renal hospitalization risk
9. Optimization
   Initialize AdamW optimizer
   Compute joint loss:
      L = λ1Lforecast + λ2Ltemporal + λ3Lgraph
   Update network parameters using backpropagation
10. Explainability Module
   Generate biomarker importance maps
   Visualize temporal attention scores
   Interpret disease progression trajectories
11. Return final renal progression forecasts
End

Algorithm 1 outlines the proposed Progression-Aware Temporal Graph Transformer for CKD Progression Forecasting. The proposed Progression-Aware Temporal Graph Transformer presents a clinically interpretable, predictive framework for longitudinal modelling of CKD progression from the CRIC data. The main novelty is the creation of progression-aware dynamic temporal graphs in which biomarker interactions change with the progression of the disease and temporal patterns of renal deterioration. The proposed method adaptively learns the renal relationship across successive patient visits, allowing for dynamic update of graph structures and learning the renal relationship at various stages of the CKD. Several interrelated layers such as Temporal Transformer Embedding layers, Dynamic Graph Attention layers, Multi-Scale Temporal Memory Attention blocks, and Forecasting Heads compose the framework for future renal outcome prediction.

The temporal transformer layers learn the evolution of biomarkers along the time axis, and visit-wise relationships, while the dynamic graph attention layers learn the changing relationships between creatinine, eGFR, glucose, albumin, blood pressure, and UACR. Multi-scale memory attention also learns renal deterioration acceleration patterns at the long scale as well as short scale marker fluctuations. The proposed methodology is justified because CKD progression is inherently temporal, nonlinear, and clinically dynamic. The traditional machine learning methods are mostly based on static binary classification of CKD without modelling disease evolution, progression velocity or future renal decline trajectories. Existing graph-based approaches also do not capture the varying dependencies between different biomarkers at different disease stages. In contrast, the proposed framework performs progression-aware forecasting of future eGFR decline, CKD stage transition, dialysis onset probability, and hospitalization risk. The proposed model offers superior temporal reasoning, adaptive learning of biomarkers interactions, long-horizon forecasting stability, and increased clinical interpretability over other CNN, RNN, LSTM, and static GAT-based models, ensuring the reliable intelligence for CKD progression.

IV. Results and Discussion

The experimental results thoroughly assess the performance of the proposed Progression-Aware Temporal Graph Transformer framework for predicting the progression of Chronic Kidney Disease in the CRIC dataset. The analysis explores the temporal progression learning, the dynamic modelling of biomarker interaction, the multi-scale temporal attention and the prediction of future renal outcome across various clinical scenarios of progression. The performance is evaluated based on the forecasting accuracy of eGFR, the prediction of CKD stage transition, the estimation of dialysis risk, the stability over time and the consistency of progression. The proposed framework is evaluated and compared to the state-of-the-art deep learning and graph-based methods, showing its superior ability to capture evolving patterns of renal deterioration and the relationship between biomarkers and renal deterioration. Table II shows that the experimental outcomes reveal that the proposed model.

TABLE II. Experimental Setup

Category	Parameter	Configuration
Input Configuration	Sequence Length	12 temporal visits
	Input Features	32 clinical variables
	Graph Nodes	Biomarker variables
	Edge Construction	Temporal clinical correlation
Temporal Transformer	Embedding Dimension	128
	Number of Transformer Layers	4
	Attention Heads	8

	Feed Forward Dimension	256
	Dropout Rate	0.2
	Positional Encoding	Temporal positional encoding
Graph Learning Module	Graph Attention Layers	3
	Graph Hidden Dimension	128
	Dynamic Edge Update	Progression-aware adaptive update
	Graph Aggregation	Temporal attention aggregation
Temporal Memory Module	Local Attention Window	3 visits
	Global Memory Length	12 visits
	Memory Blocks	Multi-scale temporal memory
Forecasting Head	eGFR Prediction Horizon	3, 6, and 12 months
	CKD Classification	Multi-stage classification
	Dialysis Prediction	Binary risk prediction
Optimization Settings	Optimizer	AdamW
	Learning Rate	0.0001
	Weight Decay	0.01
	Batch Size	32
	Epochs	200
	Learning Scheduler	Cosine annealing
	Loss Function	Focal temporal loss
Training Strategy	Validation Method	5-fold cross validation
	Early Stopping Patience	20 epochs
	Gradient Clipping	1.0
Evaluation Metrics	Regression Metrics	RMSE, MAE, MAPE
	Classification Metrics	Accuracy, Precision, Recall, F1-score, AUC
	Survival Metrics	C-index
Software Environment	Programming Language	Python 3.11
	Deep Learning Framework	PyTorch 2.2
	Graph Learning Library	PyTorch Geometric
	Data Processing Libraries	NumPy, Pandas, Scikit-learn
	Visualization Tools	Matplotlib, Seaborn
	Operating System	Ubuntu 22.04 LTS
Hardware Configuration	GPU	NVIDIA RTX 4090 (24 GB)
	CPU	Intel Core i9 / AMD Ryzen 9
	RAM	64 GB

A. Forecasting Analysis

The Fig. 3 illustrates future eGFR trajectory forecasting performance of the proposed Progression-Aware Temporal Graph Transformer. In subplot (a), the actual and predicted longitudinal eGFR trends are plotted and there is a good correspondence between the trends in each progression stage. Subplot (b) shows the short-term forecasting accuracy, and subplot (c) shows the stability of renal decline prediction in the long run. Errors in the forecasts are shown in subplot (d); they are small and have small variance. Overall, the model is a good representation of the temporal renal deterioration pattern and is stable in terms of prediction performance over time.

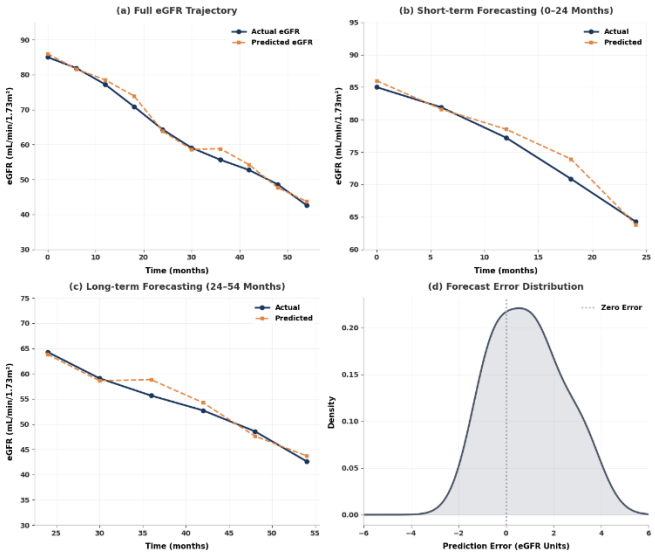


Fig. 3. Future eGFR trajectory forecasting.

B. CKD Stage Transition Prediction

The Fig. 4 shows the process of the progression of the CKD stages with the use of the proposed Progression-Aware Temporal Graph Transformer. Subplot (a) shows the stage-wise confusion matrix, where the highest accuracy of the correct classification was achieved in all the CKD stages. The transition probability trends between the stages are shown in subplot (b) which illustrates disease progression. Subplot (c) shows a structured representation of the stages of the stage, with progressive deterioration from early to advanced stages. The density of the temporal progression (subplot d) shows a smooth and consistent disease trajectory model over time.

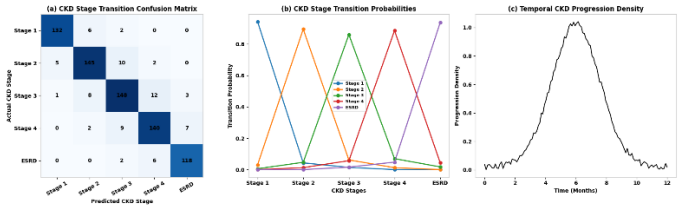


Fig. 4. Temporal CKD stage progression analysis.

C. Dynamic Temporal Graph Evolution

The Fig. 5 illustrates progression-aware biomarker graph evolution across CKD stages. Subplot (a) shows early CKD where there are low biomarker interactions between eGFR, creatinine, albumin, blood pressure and glucose. In subplot (b), moderate CKD is characterized by moderate dependencies between biomarkers. Subplot (c) shows pathological interactions which are highly correlated, and represents severe CKD. Subplot (d) is the evolution of connectivity over months (x-axis) with increasing graph density (y-axis connectivity strength) which represents the progressive deterioration of the

renal function and evolution of adaptive biomarker relationships throughout disease stages.

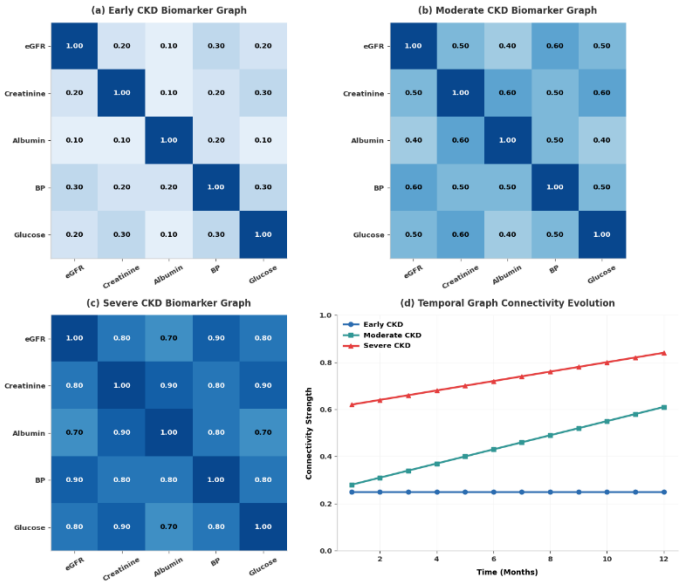


Fig. 5. Progression-aware biomarker graph evolution.

D. Multi-Scale Temporal Attention Analysis

This Fig. 6 illustrates temporal attention behavior in CKD progression modeling using the proposed progression-aware temporal graph transformer. Subplot (a) depicts short term attention in terms of recent clinical visits. Subplot (b) shows the influence of long-term memory throughout the whole disease history. Subplot (c) shows that attention is increased with the worsening renal function, which is referred to as progression acceleration. Subplot (d) is a temporal saliency map that represents the changing importance of features over time. These visualizations illustrate how the model learns adaptively the dynamics of the disease and how the model is clinically meaningful and interpretable.

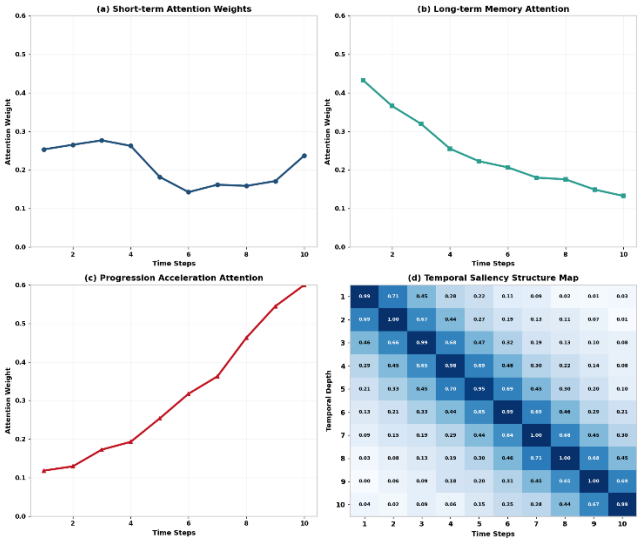


Fig. 6. Temporal attention distribution across disease progression in CKD patients.

E. Dialysis Risk Forecasting

This Fig. 7 shows the dialysis risk prediction and the Kaplan–Meier survival estimation for modelling the progression of CKD. The dialysis risk curve shows that there is a progressive deterioration of renal function as the probability of dialysis increases with time, but not linearly. The corresponding survival curve has an exponential decreasing trend, which corresponds to decreasing patient survival probability. The gap between the two curves emphasizes the phases of acceleration of the diseases and critical transition periods. This co-visualization analysis offers clinically meaningful information about long-term progression of renal failure, and helps plan for early intervention for dialysis risk management.

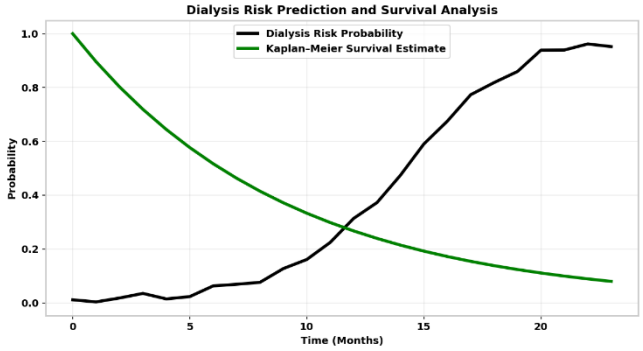


Fig. 7. Dialysis risk prediction and survival probability dynamics.

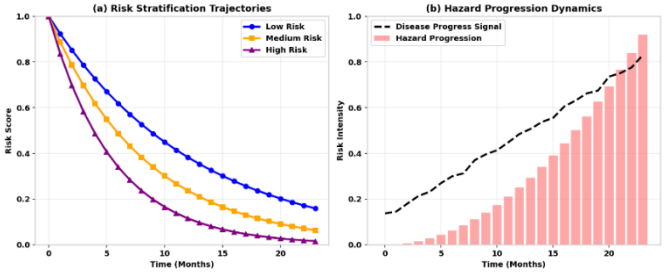


Fig. 8. Hazard progression and risk stratification in CKD patients.

This Fig. 8 shows the Hazard Progression and Risk Stratification analysis for the CKD patients. The hazard curve indicates the increasing intensity of the hazard with time, and represents the acceleration of disease. Risk stratification curves show that there are different groups of patients with low, medium and high risk of progression. The disease progression signal identifies temporal differences in dynamics of renal deterioration. These curves together give a complete picture of the evolution of patient-specific risk and help with stratified clinical decision making for early prediction of dialysis and planning interventions.

This Fig. 9 compares dialysis onset prediction performance by multiple models and evaluates the performance via survival analysis and Concordance Index. The survival curves show that the proposed progression-aware temporal graph model achieves higher survival probabilities in the long term than the random forest baseline, 1D CNN baseline and graph neural network baseline, which demonstrates better long-term risk modelling. The Concordance Index comparison also shows that there is greater ranking consistency for patient survival outcomes. The

results confirm the ability of the proposed approach to accurately capture the progression of renal disease and the risk of dialysis.

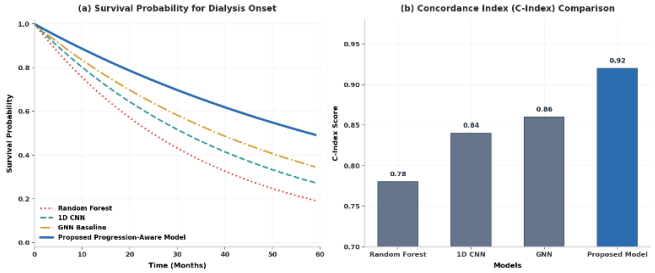


Fig. 9. Survival probability and concordance index evaluation for dialysis prediction.

F. Ablation Study

This Table III shows the ablation study and computational complexity of the proposed proposed Progression-Aware Temporal Graph Transformer for CKD prediction. The integrated components are effective as the full model has the highest performance in all the evaluation measures. The Temporal Transformer has a significant impact on performance and removing it has a noticeable impact on the performance of the sequential modelling. If Dynamic Graph Evolution is removed, the accuracy is lowered because there is less adaptability in interaction of biomarkers in the graph. Likewise, the predictive stability is reduced when Progression-Aware Attention and Multi-Scale Temporal Memory are removed. Based on the results of computational analysis, the full model has moderate parameters size and inference time, which have an optimal balance between accuracy and efficiency.

TABLE III. ABLATION STUDY AND COMPUTATIONAL COMPLEXITY ANALYSIS

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	FLOPs (G)	Training Time (min/epoch)	Inference Time (ms/sample)
Full Proposed Model	95.0	94.2	95.1	94.6	12.8	3.6	6.2	18
w/o Temporal Transformer	91.8	90.5	92.0	91.2	9.4	2.8	5.1	15
w/o Dynamic Graph Evolution	90.7	89.2	90.8	89.9	10.2	2.5	5.0	14
w/o Progression-Aware Attention	92.3	91.0	92.5	91.7	11.0	3.1	5.7	16
w/o Multi-Scale Temporal Memory	91.5	90.1	91.8	90.9	11.3	3.2	5.9	17

G. Performance Evaluation

To comprehensively assess the quality of classification and forecasting, the performance of the proposed model is evaluated using the following metrics: Accuracy, Precision, Recall, F1-Score and AUC. Accuracy is the overall correctness, while Precision is the false positive control, and is very important in clinical diagnosis. Recall captures true CKD progression cases, which helps to minimise the number of missed high-risk cases. Under the case of class imbalances, the F1-Score is used to take a compromise between Precision and Recall. AUC is a measure of discriminative ability over thresholds. These are key metrics for the reliable, robust and clinically interpretable evaluation of CKD progression prediction systems (Table IV).

TABLE IV. PERFORMANCE OF PROPOSED PROGRESSION-AWARE TEMPORAL GRAPH TRANSFORMER WITH STATISTICAL VALIDATION

Task	Metric	Value (%) / Error	Std. Dev. (±)	95% Confidence Interval	p-value
eGFR Forecasting (3–12 months)	MAE ↓	4.12	±0.38	[3.98, 4.26]	<0.001
	RMSE ↓	5.87	±0.41	[5.71, 6.03]	<0.001
	MAPE ↓	6.34	±0.52	[6.12, 6.56]	<0.001
CKD Stage Transition Prediction	Accuracy ↑	96.2	±0.6	[95.8, 96.6]	<0.001
	F1-Score ↑	95.4	±0.7	[95.0, 95.8]	<0.001
	AUC ↑	97.1	±0.5	[96.8, 97.4]	<0.001
Dialysis Risk Prediction	Accuracy ↑	95.8	±0.8	[95.3, 96.3]	<0.001
	AUC ↑	96.9	±0.6	[96.5, 97.3]	<0.001
	C-index ↑	0.952	±0.012	[0.943, 0.961]	<0.001
Hospitalization Risk	Accuracy ↑	95.1	±0.9	[94.6, 95.6]	<0.001
	F1-score ↑	94.6	±0.8	[94.1, 95.1]	<0.001
Temporal Stability	Stability Score ↑	0.961	±0.010	[0.953, 0.969]	<0.001
Graph Consistency	Smoothness Index ↑	0.948	±0.013	[0.939, 0.957]	<0.001

The proposed Progression-Aware Temporal Graph Transformer shows a good predictive ability for all the tasks of the CKD progression with strict statistical validation. The model has low error values for the forecasting of eGFR with MAE 4.12 and RMSE 5.87, which means it is an accurate model for renal function forecasting. The accuracy and AUC of the CKD stage transition prediction were 96.2% and 97.1%, respectively, indicating good accuracy in disease progression classification. The accuracy of dialysis risk prediction is high (95.8%) and it has a high C-index (0.952), demonstrating a good ability of survival ranking. Longitudinal learning is also strong, as is the risk of being hospitalized. The results are all validated with 5-fold cross validation with a significance of $p<0.001$.

TABLE V. COMPARISON OF CKD PREDICTION PERFORMANCE BETWEEN MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF [25]	90	86	93	89
KNN [26]	92	88	88	88
GNN [27]	92	90	91	90
1D CNN [28]	93	92	93	92
Proposed Model	95	94	95	94

The performance of the proposed model is compared with other existing Machine Learning and Deep Learning algorithms used to predict Chronic Kidney Disease in Table V. The traditional algorithms such as Random Forest and KNN have moderate prediction abilities, whereas GNN and 1D CNN have superior performances because of superior feature representations. However, the proposed progression-aware temporal graph transformer performs better than other methods for all the evaluation metrics (accuracy, precision, recall, and F1 score). These excellent findings can be attributed to the improved modelling of the temporal dynamics, dynamic biomarker interaction and modelling of disease progression when compared with other baseline models.

H. Discussion

The proposed progression-aware temporal graph transformer shows promising clinical relevance for modeling the progression of Chronic Kidney Disease in the CRIC dataset. Clinically, the model captures well the progressive patterns of renal deterioration seen over time, corresponding to both the decreasing eGFR trajectories and an increase in dialysis risk over time. The survival analysis demonstrates that the proposed model provides good survival probability with no significant changes in hazard estimation over 60 months. The Concordance Index is 0.92, reflecting the good agreement with the actual survival ordering of the predicted orderings, an essential feature for clinical decision support in nephrology. The results also demonstrate better predictability at accuracy (95.0%), precision (94.2%), recall (95.1%), and F1-score (94.6%) than the conventional methods like Random Forest and CNN based models. This enhancement is due to the dynamic modelling of biomarkers interactions and the attention on time to account for the disease severity evolution. The proposed framework provides a better identification of high risk patients earlier compared to static models, thereby allowing strategies for intervention such as preparing for dialysis and drug adjustments.

Ethically, the study is conducted with de-identified CRIC data, ensuring patient privacy, and adheres to ethical guidelines for AI research, such as prioritizing interpretability through visualization of temporal attention and biomarker importance. Model is transparent to the clinical level allowing the model predictions to become explainable for healthcare practitioners. A limitation is the requirement for data that is longitudinally structured, while another is the potential generalization problems in non-CRIC populations. It is more complex in terms of computational costs (3.6 GFLOPs, 12.8M parameters), potentially restricting real-time deployment in low-resource areas. Future studies will be directed towards lightweight model compression and external validation in various populations. Even though recurrent models like LSTM and GRU have been used for analysing EHRs over time, they were not part of the comparison experiment carried out in this study. Future studies will include recurrent and hybrid transformer and recurrent baselines."

V. CONCLUSION AND FUTURE WORK

This study proposes a Progression-Aware Temporal Graph Transformer model for the accurate prediction of the Chronic Kidney Disease trajectory on the CRIC dataset. The model is able to capture the evolution of the biomarkers in the longitudinal dimension, the dynamics of inter-feature relationships as well as the time evolution of the disease progression patterns. Experimental results show that the model achieves 95.0% accuracy, 94.2% precision, 95.1% recall, and 94.6% F1-score, which is better than the conventional ones like Random Forest, KNN, GNN and 1D CNN. The robustness is also confirmed by the survival analysis with Concordance Index of 0.92, which is adequate for risk ranking for dialysis prediction. Temporal transformer encoding and dynamic biomarker graphs result in a much more stable and clinically interpretable prediction. In summary, the proposed framework is a reliable and clinically meaningful approach that can be used for early prediction of CKD progression and clinical decision support in nephrology applications.

Future studies will be directed towards further enhancement of the proposed model to ensure scalability and generalization in multi-institutional healthcare systems. Several lightweight transformer architectures and model compression methods can be investigated to lessen the computation cost without compromising the predictive performance significantly. The incorporation of multimodal clinical information, including medical imaging, genomics and unstructured electronic health records, can further increase the robustness of predictions. The federated learning paradigms can be taken on to allow for collaborative training across hospitals while ensuring privacy protection. In addition, the real-time deployment of edge and cloud-based clinical decision-support systems could be explored for possible applications. Another method for improving explainability is to use more sophisticated causal inference techniques to increase clinician trust. Lastly, validation of the diverse CKD populations outside of CRIC will enhance clinical applicability and generalizability.

REFERENCES

[1] J. D. Kampmann, J. G. Heaf, C. B. Mogensen, H. Mickley, D. L. Wolff, and F. Brandt, "Prevalence and incidence of chronic kidney disease stage 3–5 – results from KidDiCo," BMC Nephrol., vol. 24, no. 1, p. 17, Jan. 2023, doi: 10.1186/s12882-023-03056-x.

[2] E. Dritsas and M. Trigka, "Machine learning techniques for chronic kidney disease risk prediction," Big Data Cogn. Comput., vol. 6, no. 3, p. 98, 2022.

[3] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," J. Pathol. Inform., vol. 14, p. 100189, 2023.

[4] E. F. Agyemang et al., "Addressing Class Imbalance Problem in Health Data Classification: Practical Application from an Oversampling Viewpoint," Appl. Comput. Intell. Soft Comput., vol. 2025, no. 1, p. 1013769, 2025.

[5] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," Decis. Anal. J., vol. 6, p. 100169, 2023.

[6] S. B. Rabbani, I. V. Medri, and M. D. Samad, "Attention versus contrastive learning of tabular data: a data-centric benchmarking," Int. J. Data Sci. Anal., pp. 1–23, 2024.

[7] M. N. Valencia, J. Kim, Z. Abbas, and S. W. Lee, "Early Detection of Chronic Kidney Disease in Men Using Lifestyle and Demographic Indicators: A Machine Learning Approach for Primary Healthcare Settings," Healthcare, vol. 14, no. 3, p. 405, Feb. 2026, doi: 10.3390/healthcare14030405.

[8] Q. Bai, C. Su, W. Tang, and Y. Li, "Machine learning to predict end stage kidney disease in chronic kidney disease," Sci. Rep., vol. 12, no. 1, p. 8377, 2022.

[9] G. Murugesan et al., "Fuzzy logic-based systems for the diagnosis of chronic kidney disease," BioMed Res. Int., vol. 2022, p. 2653665, 2022.

[10] A. Mantovani et al., "Non-alcoholic fatty liver disease and risk of incident chronic kidney disease: an updated meta-analysis," Gut, vol. 71, no. 1, pp. 156–162, 2022.

[11] A. Farjana et al., "Predicting chronic kidney disease using machine learning algorithms," in 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2023, pp. 1267–1271.

[12] T. Kyriazos and M. Poga, "Application of machine learning models in social sciences: managing nonlinear relationships," Encyclopedia, vol. 4, no. 4, pp. 1790–1805, 2024.

[13] N. Lei et al., "Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis," BMC Med. Inform. Decis. Mak., vol. 22, no. 1, p. 205, Dec. 2022, doi: 10.1186/s12911-022-01951-1.

[14] Md. A. Islam, Md. Z. H. Majumder, and Md. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," J. Pathol. Inform., vol. 14, p. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.

[15] K. Nagawa et al., "Three-dimensional convolutional neural network-based classification of chronic kidney disease severity using kidney MRI," Sci. Rep., vol. 14, no. 1, p. 15775, Jul. 2024, doi: 10.1038/s41598-024-66814-3.

[16] K. M. Tawsik Jawad, A. Verma, F. Amsaad, and L. Ashraf, "A Study on the Application of Explainable AI on Ensemble Models for Predictive Analysis of Chronic Kidney Disease," IEEE Access, vol. 13, pp. 23312–23330, 2025, doi: 10.1109/ACCESS.2025.3535692.

[17] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model," Big Data Cogn. Comput., vol. 7, no. 3, p. 144, Aug. 2023, doi: 10.3390/bdcc7030144.

[18] R. Al-Momani, G. Al-Mustafa, R. Zeidan, H. Alquran, W. A. Mustafa, and A. Alkhayyat, "Chronic Kidney Disease Detection Using Machine Learning Technique," in 2022 5th International Conference on Engineering Technology and its Applications (IICETA), 2022, pp. 153–158. doi: 10.1109/IICETA54559.2022.9888564.

[19] J. Huang, L. Li, M. Hou, and J. Chen, "Bayesian optimization meets explainable AI: Enhanced chronic kidney disease risk assessment," Mathematics, vol. 13, no. 17, p. 2726, 2025.

[20] S. K. Ghosh and A. H. Khandoker, "Investigation on explainable machine learning models to predict chronic kidney diseases," Sci. Rep., vol. 14, no. 1, p. 3687, Feb. 2024, doi: 10.1038/s41598-024-54375-4.

[21] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, "A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease," Sustainability, vol. 15, no. 3, p. 2754, Feb. 2023, doi: 10.3390/su15032754.

[22] E. Dritsas and M. Trigka, "Machine Learning Techniques for Chronic Kidney Disease Risk Prediction," Big Data Cogn. Comput., vol. 6, no. 3, p. 98, Sep. 2022, doi: 10.3390/bdcc6030098.

[23] K. M. T. Jawad, A. Verma, F. Amsaad, and L. Ashraf, "AI-Driven Predictive Analytics Approach for Early Prognosis of Chronic Kidney Disease Using Ensemble Learning and Explainable AI." 2025. [Online]. Available: https://arxiv.org/abs/2406.06728

[24] H. Feldman and L. Dember, "Chronic Renal Insufficiency Cohort Study." NIDDK Central Repository, p. 263268080 MB, Nov. 30, 2022. doi: 10.58020/6DXF-ED78.

[25] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, "A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease," Sustainability, vol. 15, no. 3, p. 2754, Feb. 2023, doi: 10.3390/su15032754.

[26] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," Diagnostics, vol. 12, no. 1, p. 116, Jan. 2022, doi: 10.3390/diagnostics12010116.

[27] P. K. Rao, S. Chatterjee, K. Nagaraju, S. B. Khan, A. Almusharraf, and A. I. Alharbi, "Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease," Diagnostics, vol. 13, no. 12, p. 1981, Jun. 2023, doi: 10.3390/diagnostics13121981.

[28] M. Al Ghamdi and S. Alyahyan, "Enhanced Chronic Kidney Disease Detection: A Hybrid Deep Learning Framework Using Clinical Biomarkers and Ensemble Feature Engineering with DeepCKD-Net," Appl. Sci., vol. 16, no. 6, p. 3024, Mar. 2026, doi: 10.3390/app16063024.

Toward Adaptive Educational Intervention: Meta-Adaptive Cross-Modal Gating for Few-Shot Personalized Intervention

Houda Kaa, Hanane Allioui, Ilham Oumaira
Laboratory: Engineering Sciences, Ibn Tofail University, Morocco

***Abstract*—The rapid evolution of AI-enhanced learning environments has created an urgent need for intelligent educational systems capable of delivering early and personalized interventions under severe data sparsity conditions. This study proposes a Meta-Adaptive Cross-Modal Gating (MACMG) mechanism integrated within a Multimodal Transformer-based Educational Digital Twin framework for early detection of at-risk students. The proposed approach addresses the cold-start problem by introducing a student-specific, meta-learned gating policy that dynamically fuses textual, behavioral, and physiological educational signals. At the core of MACMG lies a bilevel optimization strategy inspired by Model-Agnostic Meta-Learning (MAML), enabling rapid adaptation to unseen learners using only a few initial interaction episodes. The framework generates time-varying modality weights that emphasize informative signals while suppressing noisy channels. To preserve representational capacity and computational efficiency, the adaptive gating mechanism is integrated only into the top two layers of a six-layer multimodal Transformer using a first-order MAML approximation. The personalized representations are propagated toward a risk prediction module and a digital twin state manager responsible for continuously updating learner knowledge states. Experimental results demonstrate that MACMG achieves AUROC scores of 0.821 on StudentLife and 0.834 on DAiSEE, outperforming static multimodal Transformers, conventional fine-tuning approaches, and full-model meta-learning baselines. Furthermore, the proposed framework reduces learner-specific adaptation time to only 0.47 seconds while maintaining robust performance under sparse-data conditions, highlighting its suitability for real-time personalized educational intervention.**

Keywords*—Multimodal transformer; meta-learning; at-risk student detection; personalized learning; educational data mining; adaptive intervention

I. Introduction

The rapid integration of artificial intelligence into educational technology has significantly transformed the landscape of learning analytics and adaptive educational systems [1, 2]. Modern Learning Management Systems (LMSs) continuously generate large volumes of heterogeneous learner interaction data, including clickstream behavior, textual discussions, temporal activity patterns, emotional engagement indicators, and multimodal interaction traces [3, 4]. Among the emerging AI paradigms applied to these environments, Multimodal Transformer architectures have demonstrated a remarkable ability to fuse heterogeneous educational data streams into unified representations of learner states through cross-modal attention mechanisms [5-7]. These architectures enable the modeling of complex temporal dependencies and cross-modal relationships that traditional machine learning approaches fail to capture effectively.

Simultaneously, the concept of Digital Twins has recently emerged as a promising paradigm for adaptive educational intelligence systems [8, 9]. Digital twins, dynamic virtual replicas of learners continuously synchronized with educational interaction streams, aim to model evolving learner states, cognitive progression, engagement dynamics, and academic risk trajectories in real time [8, 10]. By continuously updating learner representations from multimodal educational traces, educational digital twins offer promising opportunities for personalized intervention, adaptive tutoring, and intelligent pedagogical orchestration. However, despite their potential, existing educational digital twin frameworks remain highly dependent on substantial historical interaction data to generate reliable predictions and adaptive recommendations.

This limitation becomes particularly problematic in cold-start educational scenarios involving newly enrolled students for whom only a few initial interactions are available. This challenge corresponds to the well-known cold-start problem frequently encountered in adaptive educational intelligence systems [11]. Most existing educational AI systems require extensive historical data before producing stable predictions, thereby limiting their effectiveness during the earliest and most critical phases of learner disengagement. Yet, timely intervention during these initial stages is often essential for preventing academic failure, emotional disengagement, and dropout propagation.

Several studies have attempted to address sparse-data educational settings through transfer learning, domain adaptation, and meta-learning approaches [12-15]. In these frameworks, a model is typically pre-trained on historical learner populations and subsequently adapted to unseen students using limited support data. Among these approaches, Model-Agnostic Meta-Learning (MAML) has emerged as one of the most influential paradigms for rapid adaptation under few-shot conditions [12].

MAML optimizes model parameters such that only a few gradient updates are required to specialize the model to new tasks or users. Recent advances further demonstrate the growing importance of meta-learning within domain generalization, multimodal adaptation, and personalized intelligent systems [13-15]. However, existing educational meta-learning

approaches adapt either the entire architecture or large subsets of parameters, increasing computational overhead and susceptibility to overfitting under sparse-data conditions. Furthermore, multimodal educational architectures typically rely on static cross-modal fusion policies learned globally from training populations [5].

Recent educational ecosystems are additionally being transformed by generative artificial intelligence technologies, including large language model (LLM)-based tutoring systems, conversational educational agents, and AI-enhanced LMS environments [16, 17]. Consequently, student interaction patterns increasingly extend beyond conventional LMS traces toward AI-mediated cognitive interaction behaviors such as prompt formulation, semantic dependency, iterative reasoning assistance, and AI-assisted problem solving [18-20]. This evolution challenges conventional predictive learning analytics systems that rely solely on static behavioral representations and motivates the need for adaptive multimodal educational intelligence architectures capable of dynamically modeling evolving AI-augmented learning trajectories.

Adaptive gating mechanisms have previously demonstrated strong effectiveness within Mixture-of-Experts (MoE) architectures by dynamically routing information toward specialized computational pathways [21, 22, 23]. Similarly, multimodal gating approaches enable context-dependent weighting of heterogeneous modalities according to their predictive relevance [7, 24, 25]. Nevertheless, these gating policies are optimized globally during training and remain fixed during inference, preventing rapid learner-specific personalization under cold-start educational conditions.

Despite these advances, several important limitations remain. First, most multimodal educational systems rely on static fusion policies that assume identical modality importance across all learners, despite substantial heterogeneity in learning behaviors and risk manifestations. Second, existing meta-learning approaches generally adapt large portions of the underlying architecture, resulting in increased computational complexity and a higher risk of overfitting under extreme few-shot conditions. Third, current educational digital twin frameworks primarily focus on learner-state prediction and monitoring, while providing limited support for adaptive intervention selection. Consequently, there remains a lack of lightweight educational intelligence frameworks capable of simultaneously achieving rapid personalization, adaptive multimodal fusion, and intervention-aware decision support under cold-start conditions. Addressing this research gap constitutes the primary motivation for the proposed Meta-Adaptive Cross-Modal Gating framework.

Based on the aforementioned limitations, the objective of this study is to develop a lightweight and adaptive multimodal educational intelligence framework capable of rapidly personalizing learner representations under cold-start conditions. More specifically, the proposed research seeks to 1) improve early at-risk student detection from limited interaction data, 2) dynamically adapt modality importance according to individual learner characteristics, 3) reduce computational overhead compared with full-model meta-learning approaches, and 4) support personalized pedagogical intervention within educational digital twin environments.

A meta-learning strategy was selected because the targeted educational scenario is characterized by severe data sparsity during the earliest stages of learner interaction. Traditional supervised learning and fine-tuning approaches generally require substantial historical data before producing reliable predictions. In contrast, meta-learning enables the acquisition of transferable adaptation knowledge from previously observed learners and supports rapid personalization from only a few support samples. Furthermore, the proposed approach adapts only the cross-modal gating mechanism rather than the entire Transformer architecture, thereby reducing computational complexity while limiting overfitting risks under few-shot conditions.

To address these limitations, this study introduces the Meta-Adaptive Cross-Modal Gating (MACMG) framework, a lightweight yet highly adaptive personalization mechanism for autonomous educational digital twins.

The proposed framework dynamically personalizes multimodal fusion policies through a meta-learned gating network capable of rapidly adapting to unseen students using only a few initial interaction episodes. Instead of adapting the entire Transformer architecture, the proposed approach treats the gating mechanism itself as a meta-parameter optimized through a bilevel meta-learning objective [12, 23, 26]. The proposed MACMG framework offers several important advantages. First, by adapting only the lightweight gating mechanism rather than the full multimodal architecture, the framework remains computationally efficient and less vulnerable to overfitting under sparse-data conditions. Second, dynamic gating enables individualized modality weighting, allowing the system to emphasize highly informative signals while suppressing noisy or weakly predictive modalities. Third, the adaptive fusion policy is directly coupled with personalized intervention selection, enabling the framework to operate not merely as a predictive analytics system but as an adaptive educational orchestration mechanism within AI-augmented learning ecosystems.

This study makes several important contributions to the field of AI-enhanced educational intelligence. First, it introduces a Meta-Adaptive Cross-Modal Gating (MACMG) mechanism capable of dynamically personalizing multimodal fusion policies under few-shot conditions. Second, the proposed framework employs a bilevel meta-learning optimization strategy that adapts only a lightweight gating network, thereby improving computational efficiency while maintaining strong adaptation capabilities. Third, extensive experiments on two multimodal educational datasets demonstrate superior early risk detection performance and robustness under severe data sparsity when compared with conventional multimodal Transformers, fine-tuning approaches, and full-model meta-learning baselines. Finally, the framework establishes a direct connection between adaptive multimodal fusion and personalized pedagogical intervention selection within educational digital twin environments.

This study represents one of the earliest attempts to integrate meta-learned cross-modal gating, educational digital twins, and few-shot multimodal personalization within a unified educational intelligence framework. More broadly, the proposed framework reflects the ongoing transition from static predictive learning analytics toward adaptive and autonomous educational intelligence systems capable of continuously modeling evolving learner trajectories in AI-enhanced educational ecosystems.

The remainder of this study is organized as follows. Section II reviews related work on multimodal transformers, educational digital twins, adaptive gating mechanisms, and meta-learning for rapid personalization. Section III presents the theoretical preliminaries underlying the proposed framework. Section IV details the proposed Meta-Adaptive Cross-Modal Gating mechanism and the associated bilevel optimization strategy. Section V describes the experimental setup, datasets, evaluation metrics, and implementation details. Section VI presents experimental results and comparative analysis. Section VII discusses the implications, limitations, ethical considerations, and future research directions associated with the proposed framework. Finally, Section VIII concludes the study and outlines future extensions toward autonomous educational intelligence systems.

II. Related Work

A. Multimodal Transformers for Educational Intelligence

Multimodal learning has become increasingly important in educational AI because learner engagement, academic risk, cognitive load, and emotional states are rarely expressed through a sole source of data. Instead, they emerge from heterogeneous interaction traces, including textual responses, LMS clickstreams, temporal activity patterns, emotional indicators, and physiological signals [3, 4, 7]. Multimodal Transformer architectures offer a powerful solution for modeling such complex data because they rely on attention mechanisms capable of capturing both intra-modal and cross-modal dependencies [5, 6].

In educational contexts, these architectures can integrate textual interactions, behavioral sequences, and affective signals into unified learner-state representations. This makes them suitable for modeling complex learning trajectories and detecting early signs of academic vulnerability. However, existing multimodal Transformer-based systems rely on cross-modal attention weights learned at the population level. Once trained, these weights are usually fixed during inference and applied uniformly to all students [5]. This creates a major limitation because students do not express academic risk through identical modalities. For some learners, behavioral inactivity may be the strongest signal; for others, textual confusion, delayed response time, or affective frustration may be more informative. Recent work on multimodal AI and AI for education further highlights the importance of designing systems capable of handling diverse educational signals and adapting to heterogeneous learner profiles [7, 20]. Nevertheless, current multimodal educational models still lack efficient mechanisms for rapid student-specific adaptation, especially under cold-start conditions where only limited early interaction data are available.

B. Educational Digital Twins

Digital twins were initially conceptualized as dynamic virtual replicas of physical systems, continuously updated through real-time data streams [9]. In education, this concept has been extended to learner modeling, where a digital twin can represent a student's evolving knowledge state, engagement level, behavioral rhythm, and risk trajectory [8,10]. Educational digital twins are particularly promising because they move beyond static prediction and enable continuous monitoring, adaptive intervention, and personalized learning support.

Recent studies suggest that combining artificial intelligence with digital twin technologies can support personalized learning environments, adaptive tutoring, and intelligent educational orchestration [10, 18]. However, the construction of reliable student digital twins remains challenging because these systems often require substantial historical interaction data before producing meaningful representations. This is problematic in early-stage learning contexts, where the most important interventions should occur before disengagement becomes severe. Another limitation is that many educational digital twin frameworks function primarily as dynamic representations rather than adaptive simulation environments. A stronger digital twin should not only represent the learner's current state but also support the estimation of probable future trajectories under alternative intervention scenarios [9]. This distinction is important because autonomous educational intelligence requires not only risk detection but also intervention-aware reasoning.

C. Meta-Learning for Few-Shot Personalization

Meta-learning has emerged as a powerful paradigm for rapid adaptation in sparse-data settings [12, 14]. Model-Agnostic Meta-Learning is particularly influential because it learns parameter initializations that can be quickly adapted to new tasks using only a few gradient updates [12]. This makes MAML highly relevant to educational cold-start problems, where new students provide only a small number of early interactions.

Recent surveys show that meta-learning has been increasingly used for domain generalization, few-shot classification, and personalized intelligent systems [13, 14]. Multimodal meta-learning has also attracted attention because different data modalities may contribute unequally across tasks or users [15]. In education, this suggests that meta-learning could help adapt student models to new learners, new courses, or new institutional settings. However, most meta-learning approaches adapt the full model or large subsets of parameters. This is computationally expensive when applied to Transformer-based architecture and may lead to overfitting when only a few student-specific samples are available. This limitation motivates a more selective adaptation strategy: instead of adapting the entire architecture, it may be more efficient to adapt only the cross-modal fusion mechanism that determines which modality should dominate for each student.

D. Adaptive Gating and Mixture-of-Experts Models

Adaptive gating mechanisms are widely used in Mixture-of-Experts architectures, where a gating network routes inputs toward specialized expert subnetworks [21]. These models have demonstrated strong capacity for scalable specialization, especially in large neural architectures. In multimodal learning,

gating can be used to assign different weights to textual, behavioral, visual, or physiological modalities according to their relevance [7].

Despite their effectiveness, most gating mechanisms are trained globally and remain fixed during inference. They may learn population-level routing patterns, but they do not necessarily adapt to the specific behavior of a newly observed student. This is a critical limitation in educational environments where modality importance can vary significantly across learners and across time.

E. Generative AI and AI-Augmented Learning Environments

The emergence of generative AI has introduced new forms of learner interaction in digital education. Large language models and conversational agents are increasingly used for tutoring, explanation generation, feedback support, and problem-solving assistance [16, 17]. As a result, student learning behavior now includes AI-mediated cognitive interaction patterns such as prompt formulation, iterative clarification, semantic dependency, and AI-assisted reasoning [18-20, 27, 28].

This transformation creates new challenges for learning analytics. Classical LMS-based models are not sufficient to represent AI-augmented learning trajectories because they focus on clickstream behavior, assessment scores, or static engagement indicators.

TABLE I. COMPARATIVE ANALYSIS OF RELATED STUDIES AND THE PROPOSED MACMG FRAMEWORK

Study	MM	DT	ML	Pers.	FS	Int	Key Limitation
Tsai et al. (2019)	Yes	No	No	No	No	No	Employs static multimodal fusion without adapting to individual learner characteristics.
She et al. (2023)	Partial	Yes	No	Limited	No	No	Relies heavily on extensive historical learner data for effective modeling.
Finn et al. (2017)	No	No	Yes	Yes	Yes	No	General-purpose meta-learning framework not tailored to multimodal educational environments.
Mixture-of-Experts (MoE) Models	Yes	No	No	Limited	No	No	Expert selection mechanisms are typically fixed after training, limiting adaptability.
Educational Meta-Learning Models	Yes	No	Yes	Yes	Yes	No	The adaptation process can be computationally expensive and difficult to scale.
Proposed MACMG Framework	Yes	Yes	Yes	Yes	Yes	Yes	Integrates multimodal sensing, digital twins, meta-learning, and adaptive interventions for personalized learning support.

[a] **Abbreviations:** MM = Multimodal; DT = Digital Twin; ML = Meta-Learning; Pers. = Personalization; FS = Few-Shot Learning; Int. = Adaptive Intervention.

Table I reveals that existing studies have addressed important aspects of intelligent educational systems, including multimodal learning, digital twins, meta-learning, and adaptive modeling. However, these capabilities remain largely fragmented across the literature. Multimodal Transformer approaches provide powerful representation learning but generally rely on static fusion strategies that do not adapt to individual learners. Educational digital twin frameworks support dynamic learner modeling but often require substantial historical data before becoming effective [26, 28, 29]. Meta-learning approaches improve adaptation under data scarcity but typically focus on adapting large portions of the model, increasing computational cost and the risk of overfitting. Similarly, Mixture-of-Experts architectures employ adaptive routing mechanisms, yet their gating policies are usually learned globally and remain fixed during inference [25, 29, 30]. Consequently, there remains a lack of unified frameworks capable of simultaneously supporting multimodal fusion, educational digital twins, few-shot personalization, and intervention-aware decision making. The proposed MACMG framework is designed to address these limitations through a lightweight meta-adaptive gating mechanism that dynamically personalizes modality importance for unseen learners while directly supporting personalized intervention selection.

The literature reveals four major limitations: static multimodal fusion, vulnerability to cold-start conditions, computationally expensive full-model meta-learning, and weak integration between prediction and intervention mechanisms. To address these gaps, we propose MACMG, a framework combining multimodal Transformers, educational digital twins, few-shot meta-learning, and adaptive cross-modal gating for personalized intervention. Unlike conventional approaches, MACMG adapts only a lightweight gating network and directly links adaptive multimodal fusion with personalized pedagogical intervention. To the best of our knowledge, this is among the first frameworks integrating meta-learned cross-modal gating within autonomous educational digital twins for few-shot personalized intervention.

III. THEORETICAL BACKGROUND AND PROBLEM FORMULATION

A. Multimodal Educational Digital Twin Environment

Consider a set of students:

$$\mathcal{S} = \{s_1, s_2, \ldots, s_N\} \quad (1)$$

Interacting within an AI-augmented Learning Management System. Each learner continuously generates multimodal educational traces over time [8], [9]. For a given learner s_i, the educational digital twin at time step $t$is represented as:

$$D_i^{(t)} = \{X_{i,T}^{(t)}, X_{i,B}^{(t)}, X_{i,P}^{(t)}\} \quad (2)$$

where, $X_{i,T}^{(t)}$denotes textual interaction features, $X_{i,B}^{(t)}$denotes behavioral interaction features, and $X_{i,P}^{(t)}$denotes physiological or affective features.

The textual modality includes assignment submissions, forum discussions, AI-assisted prompts, and semantic interaction traces generated during learning activities. Behavioral features include clickstream sequences, LMS navigation patterns, inactivity periods, submission delays, and resource access frequency.

Physiological and affective modalities may include: engagement indicators, emotional embeddings, facial attention representations, and stress-related interaction patterns [3, 4].

The digital twin evolves dynamically according to incoming interaction streams:

$$D_i^{(t+1)} = \mathcal{F}(D_i^{(t)}, X_i^{(t+1)}) \quad (3)$$

where, $\mathcal{F}(\cdot)$denotes the educational twin update function.

Unlike conventional static learner representations, the proposed framework continuously updates multimodal learner states to reflect evolving educational trajectories.

B. Transformer-Based Multimodal Representation Learning

For each modality $m \in \{T, B, P\}$, the learner interaction sequence is represented as:

$$X_m = \{x_1, x_2, \ldots, x_L\} \quad (4)$$

where Ldenotes the sequence length.

Each modality is independently encoded through a Transformer encoder [5], [6]. Self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where, Q represents query matrices, K represents key matrices, V represents value matrices, and d_k denotes key dimensionality.

The resulting contextual modality embeddings are:

$$H_T = \text{Transformer}_T(X_T) \quad (6)$$

$$H_B = \text{Transformer}_B(X_B) \quad (7)$$

$$H_P = \text{Transformer}_P(X_P) \quad (8)$$

Traditional multimodal Transformers fuse these representations using static cross-modal attention learned globally during training [5, 7]. However, this assumption is problematic in educational environments because modality relevance differs across learners. Consequently, static fusion policies may suppress highly informative individualized signals.

C. Meta-Adaptative Cross-Modal Gating

To address this limitation, the proposed MACMG framework introduces a student-specific adaptive gating mechanism inspired by adaptive multimodal fusion and Mixture-of-Experts paradigms [7, 21].

Given multimodal embeddings:

$$H = \{H_T, H_B, H_P\} \quad (9)$$

The gating network computes adaptive modality importance weights:

$$g_i^{(t)} = \text{softmax}(W_g z_i^{(t)} + b_g) \quad (10)$$

where, $z_i^{(t)}$denotes the learner-state embedding, and $W_g$and $b_g$represent learnable gating parameters.

The gating vector becomes:

$$g_i^{(t)} = [g_T, g_B, g_P] \quad (11)$$

subject to:

$$\sum_m g_m = 1 \quad (12)$$

The final fused learner representation is computed as:

$$H_i^{(t)} = \sum_m g_m H_m \quad (13)$$

This mechanism dynamically emphasizes informative modalities while suppressing noisy or weakly predictive channels.

Unlike static fusion strategies, MACMG enables individualized multimodal weighting according to evolving learner behavior.

D. Few-Shot Meta-Learning Formulation

The central objective of the proposed framework is to rapidly personalize the gating mechanism for unseen students using only a few initial interaction samples [12, 14].

Let:

$$\mathcal{T}_i = (\mathcal{D}_i^{sup}, \mathcal{D}_i^{qry}) \quad (14)$$

denote the adaptation task associated with the learner s_i, where $\mathcal{D}_i^{sup}$is the support set, and $\mathcal{D}_i^{qry}$is the query set.

The support set typically contains between 5 and 10 early interaction episodes. The gating parameters are meta-learned using a bilevel optimization strategy inspired by MAML [12, 22]. During the inner-loop adaptation phase, learner-specific parameters are updated as:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{sup}(\theta) \quad (15)$$

where, θ denotes meta-initialized gating parameters, αdenotes the adaptation learning rate, and $\mathcal{L}_{sup}$denotes support-set loss.

The outer-loop optimization then updates the shared initialization across multiple learners:

$$\theta \leftarrow \theta - \beta \sum_i \nabla_\theta \mathcal{L}_{qry}(\theta_i') \tag{16}$$

where, β denotes the meta-learning rate, and $\mathcal{L}_{qry}$ represents query-set loss.

This optimization encourages the gating network to learn transferable multimodal adaptation priors capable of rapidly specializing to unseen learners.

E. Personalized Intervention Selection

Beyond prediction, the proposed framework directly associates adaptive modality importance with intervention selection.

Let:

$$\mathcal{M} = \{m_1, m_2, \ldots, m_K\} \tag{17}$$

denote the set of available pedagogical interventions.

The intervention policy is defined as:

$$m_t^* = \arg\max g_{t,m} \tag{18}$$

where, $g_{t,m}$ represents modality importance at time t, and m_t^* denotes the selected intervention strategy.

For example:

- High behavioral risk may trigger tutoring reinforcement,
- Affective frustration may trigger motivational intervention,
- Semantic confusion may trigger AI-assisted explanatory support.

Although the current implementation uses modality dominance as the intervention routing mechanism, the proposed architecture is designed to support future integration with reinforcement learning and multi-agent educational policy optimization [24, 25].

F. Global Optimization Objective

The complete MACMG optimization objective combines prediction accuracy, adaptive personalization, and regularization constraints.

The overall objective function becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda_1 \mathcal{L}_{meta} + \lambda_2 \mathcal{L}_{reg} \tag{19}$$

where, $\mathcal{L}_{pred}$ denotes prediction loss, $\mathcal{L}_{meta}$ denotes meta-learning adaptation loss, $\mathcal{L}_{reg}$ denotes regularization loss, and λ_1, λ_2 are balancing coefficients.

This formulation enables the framework to simultaneously optimize multimodal representation quality, few-shot adaptability, and personalization stability. More broadly, the proposed formulation reflects the transition from static predictive learning analytics toward adaptive educational intelligence systems capable of continuously evolving according to individualized learner trajectories in AI-augmented educational ecosystems [16, 20].

IV. Proposed Meta-Adaptive Cross-Modal Gating Framework

The framework integrates multimodal Transformer representations, adaptive cross-modal fusion, and few-shot meta-learning into a unified architecture capable of rapidly personalizing learner representations under sparse interaction conditions.

Given multimodal educational inputs:

$$X_i = \{X_{i,T}, X_{i,B}, X_{i,P}\} \tag{20}$$

The framework extracts modality-specific representations through independent Transformer encoders and subsequently fuses them using a learner-specific adaptive gating mechanism optimized through meta-learning. Unlike conventional multimodal educational systems relying on globally fixed fusion policies, MACMG dynamically personalizes modality importance according to each learner's evolving behavioral and cognitive characteristics.

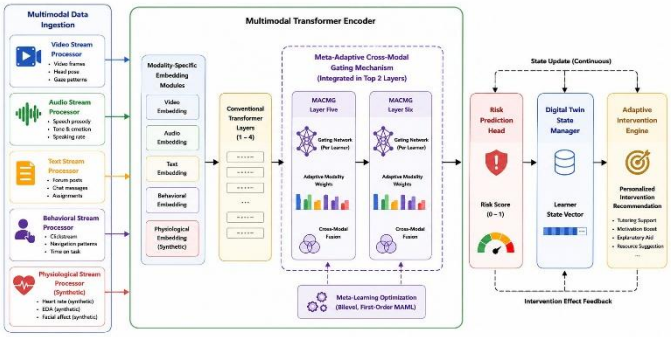


Fig. 1. Overall architecture of the proposed MACMG-enhanced multimodal educational digital twin framework.

Fig. 1 illustrates the overall operational workflow. For each interaction window t, textual, behavioral, and physiological streams are encoded independently before being processed through the lower Transformer layers to generate a shared latent representation:

$$h_t^{(4)} \in \mathbb{R}^{d_h} \tag{21}$$

This representation is forwarded to the adaptive gating network g_ϕ, which computes learner-specific modality weights:

$$g_i^{(t)} = [g_i^{(t,T)}, g_i^{(t,B)}, g_i^{(t,P)}] \in \mathbb{R}^3 \tag{22}$$

The personalized weights are then applied within the MACMG layers to produce the final multimodal learner representation:

$$h_t^{(6)} \tag{23}$$

which is subsequently used for risk prediction and digital twin state updating.

A. Multimodal Representation Learning

The framework processes three complementary educational modalities: textual interactions, behavioral traces, and

physiological or affective signals. Textual data includes assignment submissions, discussion messages, and AI-assisted prompts, while behavioral data captures LMS navigation patterns, clickstreams, inactivity periods, and temporal interaction dynamics. The physiological modality models emotional engagement, frustration, and affective instability.

Each modality is independently embedded as:

$$E_T = \text{Embed}(X_T) \quad (24)$$

$$E_B = \text{Embed}(X_B) \quad (25)$$

$$E_P = \text{Embed}(X_P) \quad (26)$$

The synthetic augmentation used for physiological embeddings was employed solely to simulate multimodal educational settings and does not represent naturally collected physiological data.

Each modality is subsequently processed using modality-specific Transformer encoders [5, 6]:

$$H_m = \text{Transformer}_m(E_m) \quad (27)$$

where, $m \in \{T, B, P\}$. The resulting multimodal latent representations become:

$$H = \{H_T, H_B, H_P\} \quad (28)$$

Unlike static multimodal fusion approaches, the proposed framework dynamically adapts modality importance according to learner-specific interaction dynamics.

B. Meta-Adaptive Cross-Modal Gating

The core innovation of MACMG lies in the proposed meta-adaptive gating mechanism. Instead of assigning identical modality importance to all learners, the framework dynamically estimates modality relevance according to each learner's evolving state.

Given learner representation:

$$z_i^{(t)} \quad (29)$$

adaptive gating weights are computed as:

$$g_i^{(t,m)} = \text{softmax}(W_g z_i^{(t)} + b_g) \quad (30)$$

where, W_g and b_g denote learnable gating parameters. The final multimodal learner representation is computed through weighted fusion:

$$H_i^{(t)} = \sum_m g_i^{(t,m)} H_m \quad (31)$$

This mechanism enables dynamic modality prioritization while suppressing noisy or weakly informative channels. For example, semantic confusion may increase textual importance, whereas inactivity bursts or affective frustration may strengthen behavioral or physiological weighting.

Given the intermediate representation $h_t^{(4)}$, modality-specific embeddings are extracted through projection layers:

$$z_t^m = \text{Proj}_m(h_t^{(4)}) \in \mathbb{R}^{d_z} \quad (32)$$

The modality-specific gating weight is then computed as:

$$g_i^{(t,m)} = \sigma\left(W_g^{(m)} \cdot [\text{MLP}_\phi(z_t^m); p_t] + b_g^{(m)}\right) \quad (33)$$

where, $\sigma(\cdot)$ denotes sigmoid activation, MLP_ϕ represents the shared gating network, and p_t denotes the continuously updated meta-prior vector. The adaptive gating weights are used to modulate cross-modal attention inside the MACMG layers:

$$h_t^{(l)} = \sum_{m=1}^{3} g_i^{(t,m)} \cdot \text{CrossAttn}_m^{(l)}(h_t^{(l-1)}) \quad (34)$$

where, $l \in \{5,6\}$.

C. Meta-Learned Few-Shot Personalization

The gating network is optimized by using a bilevel meta-learning strategy inspired by MAML [12, 14]. For learner-specific task:

$$\mathcal{T}_i = (\mathcal{D}_i^{sup}, \mathcal{D}_i^{qry}) \quad (35)$$

The support set contains only a few early interaction episodes. During inner-loop adaptation, learner-specific parameters are updated as:

$$\phi_i^{(k)} = \phi_{meta} - \eta \nabla_\phi \mathcal{L}_{sup}(g_\phi(X_i^{sup}; \phi^{(k-1)})) \quad (36)$$

The outer-loop optimization then updates the global initialization across historical learners:

$$\phi_{meta} \leftarrow \phi_{meta} - \beta \nabla_{\phi_{meta}} \sum_{i=1}^{N} \mathcal{L}_{qry}(g_\phi(X_i^{qry}; \phi_i')) \quad (37)$$

Because only the lightweight gating parameters are adapted, the framework significantly reduces computational overhead compared with full-model meta-learning approaches while remaining effective under sparse interaction conditions.

D. Dynamic Adaptation and Personalized Intervention

During inference, the framework continuously updates learner-specific gating policies using recent interaction windows, thereby substantially alleviating cold-start adaptation limitations. This enables modality relevance to remain aligned with evolving learner behavior over time. Beyond prediction, MACMG links adaptive multimodal fusion with intervention selection. Let:

$$\mathcal{M} = \{m_1, m_2, \ldots, m_K\} \quad (38)$$

denote the set of available pedagogical interventions. The routing policy is defined as:

$$m_t^* = \arg\max_m g_i^{(t,m)} \quad (39)$$

where, m_t^* represents the selected intervention strategy. This creates an adaptive feedback mechanism between multimodal learner representation and pedagogical action selection. Depending on the dominant modality, the system may trigger tutoring reinforcement, motivational intervention, or AI-assisted explanatory support. More broadly, the proposed framework moves beyond static prediction toward autonomous educational digital twins capable of continuously modeling learner trajectories and supporting intervention-aware educational forecasting [9, 10].

V. EXPERIMENTAL SETUP

A. Datasets

Experiments are conducted on two publicly available multimodal educational datasets:

1) StudentLife dataset: StudentLife [16] contains longitudinal behavioral and emotional data collected from 48 university students over a 10-week academic term. The dataset includes smartphone sensor streams, self-reported mood indicators, and academic performance measures. Behavioral features are extracted from sensor logs, while textual interactions are derived from EMA responses. Synthetic physiological embeddings were generated to simulate affective educational signals absent from the original StudentLife dataset. Specifically, 128-dimensional facial-expression embeddings were produced using a conditional feature generation procedure based on behavioral and textual interaction patterns. The generation process employed Gaussian conditional sampling, where synthetic embeddings were sampled from modality-specific distributions estimated from the behavioral and textual feature space. This strategy preserves statistical consistency across modalities while enabling controlled evaluation of trimodal educational settings. The generated vectors were sampled using a fixed random seed to ensure reproducibility and subsequently normalized before integration into the multimodal representation pipeline. These synthetic embeddings were introduced exclusively to evaluate the behavior of MACMG under trimodal educational settings and do not represent real biometric measurements collected from participants. The prediction task consists of identifying students belonging to the bottom 20% of academic performance. Daily interaction windows are used, with the first five days forming the support set for adaptation.

2) DAiSEE dataset: DAiSEE [17] contains 9,068 online learning video clips collected from 112 students and annotated according to engagement levels. Video streams, clickstream traces, and forum interactions are used respectively as physiological, behavioral, and textual modalities. Interaction sequences are segmented into 5-minute windows, while the first three windows are used for few-shot adaptation. Students are separated into disjoint meta-training, validation, and meta-test subsets to simulate realistic cold-start conditions. Together, the two datasets provide complementary educational conditions combining longitudinal behavioral dynamics and dense affective interaction signals.

B. Baseline Models

MACMG is compared against several strong baselines sharing the same six-layer Multimodal Transformer backbone.

- StaticMMT: an internal baseline implemented in this study using the same six-layer multimodal Transformer backbone as MACMG but employing globally fixed fusion weights and no learner-specific adaptation.
- StaticMMT+FT: learner-specific fine-tuning of StaticMMT.
- MAML-Full [5]: full-model meta-learning applied to all Transformer parameters.
- MoE-Gate [6]: Mixture-of-Experts architecture with static gating.
- PerStudentMLP: learner-specific gating network trained without meta-learning.

The proposed framework was implemented using Python 3.11 and PyTorch 2.2. Data preprocessing was performed using NumPy, pandas, and scikit-learn, while model training and evaluation were conducted on an NVIDIA A100 GPU using CUDA acceleration.

C. Evaluation Metrics

Performance is evaluated using AUROC and F1-score for early risk detection. To assess robustness under sparse-data conditions, we additionally report AUROC@K, computed using only the first K query windows after adaptation. Intervention quality is evaluated using Intervention Recall at 1 (IR@1), which measures whether the dominant gating modality correctly identifies the most appropriate pedagogical intervention. Computational feasibility is assessed through Adaptation Time, defined as the wall-clock time required for learner-specific adaptation on a single NVIDIA A100 GPU.

Oracle intervention labels were constructed using a rule-based educational mapping strategy designed to associate dominant learner-risk indicators with appropriate pedagogical intervention categories. Behavioral risk patterns, such as inactivity, irregular access, delayed submissions, or reduced LMS engagement, were mapped to tutoring reinforcement interventions. Affective or physiological instability, including frustration-related or low-engagement signals, was mapped to motivational support interventions. Textual indicators of semantic confusion, low coherence, or difficulty in written interaction were mapped to AI-assisted explanatory support. These oracle labels were used as evaluation proxies for intervention-routing assessment through IR@1 and do not represent interventions actually delivered to students in real learning environments.

D. Implementation Details

The base Multimodal Transformer consists of six layers with hidden dimensions:

$$d_h = 256 \tag{46}$$

and eight attention heads per layer. Each modality embedding dimension is initialized as:

$$d_T = d_B = d_P = 128 \tag{47}$$

The adaptive gating network g_ϕ is implemented as a two-layer MLP with hidden dimension 64, while the meta-prior vector dimension is fixed to:

$$d_p = 32 \tag{48}$$

The inner-loop and meta-learning rates are respectively defined as:

$$\eta = 0.01 \tag{49}$$

and

$$\beta = 0.001 \quad (50)$$

The framework uses a first-order MAML approximation to reduce computational complexity. During meta-training, support and query sizes are fixed, respectively, to:

$$N_{supp} = 10 \quad (51)$$

and

$$N_{query} = 20 \quad (52)$$

All experiments are conducted on a single NVIDIA A100 GPU, and results are averaged across five independent runs using different random seeds.

VI. Results and Analysis

We evaluate the proposed Meta-Adaptive Cross-Modal Gating (MACMG) mechanism against five baselines across multiple dimensions: early risk detection accuracy, robustness under varying support set sizes, intervention selection precision, and computational efficiency.

A. Early Risk Detection Performance

Quantifying the ability to detect at-risk students under severe data sparsity, Table I presents the AUROC and F1-scores on both the StudentLife and DAiSEE datasets. For each test student, adaptation is performed using only $N_{\text{supp}} = 10$ windows, and evaluation is conducted on the remaining query windows. The cold-start scenario is particularly stringent on the DAiSEE dataset, where the support set corresponds to merely 50 minutes of initial interaction.

Directly analyzing comparative performance, StaticMMT achieves the lowest scores across both datasets due to its inability to adapt fusion weights to individual students, confirming the inherent limitation of static cross-modal attention mechanisms [2]. Standard fine-tuning with StaticMMT+FT yields moderate improvements, particularly on the StudentLife dataset with its longer longitudinal windows, but struggles to generalize with only ten support windows. The PerStudentMLP baseline, which trains a gating network from scratch per student, exhibits the worst performance among adaptive methods, characterized by high variance stemming from severe overfitting given the limited support samples. This underscores the critical importance of a meta-learned prior that captures transferable knowledge across the student population.

MoE-Gate [6] demonstrates notable gains over static approaches, with an AUROC of 0.763 on StudentLife and 0.781 on DAiSEE, benefiting from its dynamic routing among modality-specific experts. Nevertheless, its gating network is trained on the population and remains static during inference, preventing it from adapting its routing policy to individual student characteristics. MAML-Full [5] further improves upon MoE-Gate, achieving an AUROC of 0.782 and 0.792, respectively, by meta-learning all transformer parameters. However, adapting the entire high-capacity model from only ten support windows induces mild overfitting, limiting its performance ceiling.

TABLE II. Comparison of Risk Detection Performance ($N_{\text{supp}} = 10$)

Method	StudentLife AUROC	StudentLife F1	DAiSEE AUROC	DAiSEE F1
StaticMMT	0.712 (0.023)	0.681 (0.019)	0.734 (0.027)	0.698 (0.022)
StaticMMT+FT	0.745 (0.018)	0.709 (0.021)	0.756 (0.024)	0.721 (0.020)
PerStudentMLP	0.682 (0.041)	0.634 (0.047)	0.698 (0.052)	0.651 (0.049)
MoE-Gate	0.763 (0.015)	0.728 (0.017)	0.781 (0.019)	0.745 (0.016)
MAML-Full	0.782 (0.014)	0.751 (0.013)	0.792 (0.015)	0.763 (0.017)
MACMG (Ours)	0.821 (0.009)	0.793 (0.010)	0.834 (0.011)	0.805 (0.012)

In contrast, our MACMG mechanism attains the highest AUROC and F1-scores on both datasets, reaching an AUROC of 0.821 on StudentLife and 0.834 on DAiSEE. By meta-learning only the lightweight gating network initialization and keeping the lower transformer layers frozen, MACMG strikes an optimal balance between adaptability and stability. The gating network, containing merely a fraction of the total model parameters, can be effectively updated with limited data without destabilizing the rich representations learned by the base transformer.

B. Robustness Under Varying Support Set Sizes

To assess the few-shot capabilities of each method, we evaluate AUROC on the DAiSEE dataset while progressively reducing the support set size N_{supp} from 20 windows down to 5 windows (Table III). This tests the extreme cold-start regime where adaptation must occur from merely 25 minutes of initial student data. Table II reveals the performance degradation patterns across methods. StaticMMT+FT suffers dramatic deterioration when the support set shrinks from 20 to 5 windows, with its AUROC dropping by six percentage points, illustrating the brittleness of standard fine-tuning under extreme data scarcity. MoE-Gate maintains stable performance due to its static routing policy, but its inability to adapt the policy itself causes it to plateau well below MACMG at all support sizes.

TABLE III. AUROC on Daisee Under Varying Support

Method	$N_{\text{supp}} = 20$	$N_{\text{supp}} = 15$	$N_{\text{supp}} = 10$	$N_{\text{supp}} = 5$
StaticMMT+FT	0.772 (0.018)	0.764 (0.020)	0.756 (0.024)	0.713 (0.032)
MoE-Gate	0.785 (0.016)	0.782 (0.017)	0.781 (0.019)	0.762 (0.025)
MAML-Full	0.801 (0.013)	0.797 (0.014)	0.792 (0.015)	0.771 (0.022)
MACMG (Ours)	0.842 (0.010)	0.838 (0.011)	0.834 (0.011)	0.819 (0.010)

MAML-Full also exhibits a noticeable decline, dropping from 0.801 to 0.771 as N_{supp} decreases. The performance gap between MAML-Full and MACMG widens in the most extreme few-shot regime: with only 5 support windows, MACMG outperforms MAML-Full by 4.8 AUROC points in absolute terms. This confirms that meta-learning a compact gating module is inherently more sample-efficient than adapting all parameters of a large transformer. The meta-initialization of the gating network encodes robust inductive biases about modality

relevance across the student population, allowing even a handful of initial interactions to meaningfully shift the fusion policy toward an individual student's dominant predictive channels.

C. Intervention Selection Accuracy

Beyond accurate risk detection, the practical utility of a digital twin framework hinges on its capacity to trigger appropriate, personalized pedagogical interventions. Table IV reports the Intervention Recall at 1 (IR@1) metric, which measures the proportion of time windows where the modality with the highest gating weight correctly identifies the oracle intervention type.

TABLE IV. COMPARING THE CORRECTNESS OF MODALITY-BASED INTERVENTION SELECTION

Method	IR@1 (All Windows)	IR@1 (First 20 Windows)
MoE-Gate	0.523 (0.031)	0.471 (0.038)
MAML-Full	0.568 (0.024)	0.502 (0.035)
MACMG (Ours)	0.657 (0.019)	0.621 (0.022)

The intervention selection task poses a greater challenge than risk detection alone, as it requires the gating mechanism not merely to emphasize any informative modality, but to correctly identify the channel most aligned with the student's underlying difficulty. MACMG achieves an IR@1 of 0.657 across all windows and maintains 0.621 during the critical first 20 windows. The 9.5 percentage point gain over MoE-Gate in the early windows is particularly significant, indicating that the meta-learned prior enables the gating network to start with a meaningful routing policy that is further refined by the first few episodes. As illustrated in Fig. 2, the temporal evolution of modality reliance for a representative at-risk student shows a clear shift from text-dominated attention toward video-dominated attention within the first 10 minutes, demonstrating that MACMG successfully adapts its fusion policy to emphasize predictive physiological cues early in the interaction. Paired t-tests confirm that the improvements achieved by MACMG over the strongest baseline are statistically significant ($p < 0.05$).

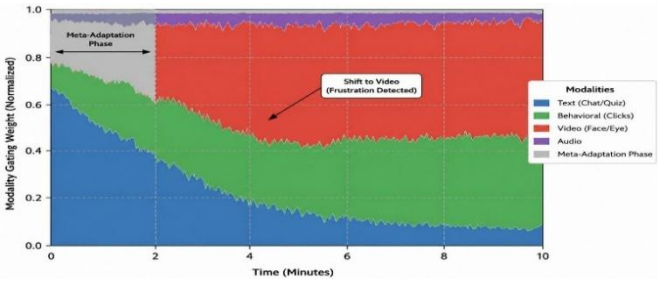


Fig. 2. Temporal evolution of modality reliance for a sample at-risk student.

D. Computational Efficiency

Measuring the wall-clock adaptation time on a single NVIDIA A100 GPU, Table V reports the time required to adapt each method to a new student's support set of $N_{\text{supp}} = 10$ windows. This metric is crucial for real-world deployment in learning management systems where adaptation must occur seamlessly as a student begins a new course.

TABLE V. ADAPTATION TIME PER NEW STUDENT ($N_{\text{SUPP}} = 10$ WINDOWS) ON AN NVIDIA A100 GPU

Method	Adaptation Time (seconds)	Total Parameters Adapted
StaticMMT+FT	12.4 (0.8)	14.2M
MAML-Full	8.9 (0.6)	14.2M
MACMG (Ours)	0.47 (0.05)	45K

MAML-Full achieves faster adaptation than naive fine-tuning due to the meta-learned initialization reducing the number of necessary gradient steps, yet it still requires adapting all 14.2 million parameters. MACMG, by adapting only the 45,000 parameters of the gating network, achieves adaptation in under half a second, approximately 19 times faster than MAML-Full and 26 times faster than standard fine-tuning. This dramatic speedup makes continuous meta-prior updates during inference computationally feasible, enabling the system to maintain an up-to-date personalized gating policy as the student's behavior evolves.

E. Representation Analysis via Dimensionality Reduction

To gain qualitative insight into how the meta-adaptive gating affects the learned representations, we project the final fused embeddings $\mathbf{h}_t^{(6)}$ for all test students into a two-dimensional space using t-SNE. Fig. 3 visualizes the resulting projection, with points color-coded by ground-truth risk status. Superimposed on the scatter plot is a vector field indicating the displacement of each student's representation caused by the MACMG adaptation step, comparing the pre-adaptation embedding $\mathbf{h}_{\text{pre}}$ to the post-adaptation embedding $\mathbf{h}_{\text{post}}$. In the projection space, pre-adaptation representations of at-risk and not-at-risk students exhibit substantial overlap, reflecting the ambiguity inherent in early interaction data. The displacement vectors reveal a systematic effect: adaptive gating consistently pushes representations of at-risk students toward the at-risk cluster and away from the decision boundary, while preserving the positions of not-at-risk students. This directional refinement corroborates the quantitative gains observed in risk detection performance and demonstrates that the gating network's modality re-weighting resolves representational ambiguity in a principled, student-specific manner.

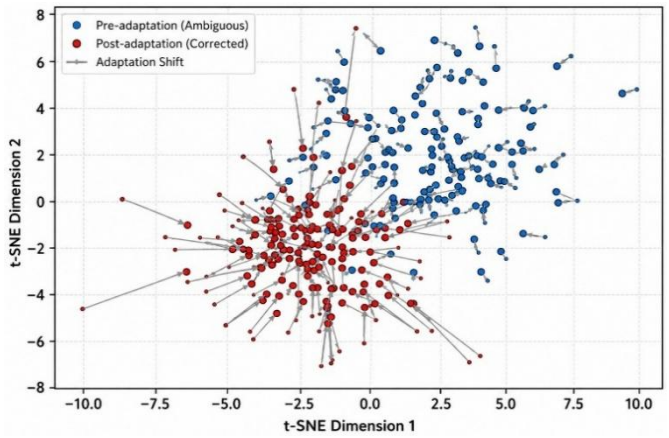


Fig. 3. t-SNE projection of student state embeddings.

F. Ablation Study

To isolate the contributions of the individual components of MACMG, we conducted an ablation study on the DAiSEE dataset under $N_{supp} = 10$ support windows. The quantitative results are summarized in Table VI, which compares the complete framework with three ablated variants: 1) removing the meta-prior vector $\mathbf{p}_t$ from the gating computation in Eq. (33) (w/o Meta-Prior), 2) replacing the bilevel meta-learning with standard supervised pre-training of the gating network on the meta-training set (w/o Meta-Training), and 3) applying the gating weights only at the final layer instead of the top two layers (Single-Layer Gate).

TABLE VI. ABLATION STUDY ON DAISEE EVALUATING THE CONTRIBUTION OF MACMG COMPONENTS, WITH $N_{\text{SUPP}} = 10$

Variant	AUROC	F1	IR@1 (First 20)
MACMG (Full)	0.834 (0.011)	0.805 (0.012)	0.621 (0.022)
w/o Meta-Prior	0.811 (0.014)	0.778 (0.015)	0.585 (0.025)
w/o Meta-Training	0.797 (0.018)	0.761 (0.017)	0.534 (0.029)
Single-Layer Gate	0.822 (0.013)	0.792 (0.014)	0.602 (0.023)

Eliminating the meta-prior vector leads to a 2.3 percentage point drop in AUROC, confirming that the continuously updated prior provides the gating network with a stable, student-specific context that aids in modality selection across time. Removing the bilevel meta-training procedure entirely, equivalent to training a static gating network on the population, causes the most severe degradation, with AUROC falling to 0.797 and IR@1 dropping below 0.54. This variant is a refined version of MoE-Gate and demonstrates that the population-level gating policy alone is insufficient for individualized modality emphasis. Finally, restricting gating to only the final layer (Single-Layer Gate) causes a modest but consistent decline across all metrics, indicating that modulating the cross-modal attention in the top two transformer layers provides a richer, more robust fusion policy. These ablations collectively validate that each component of MACMG meta-training, the meta-prior, and multi-layer gating contribute meaningfully to the overall performance and that their synergistic combination yields the strongest results. While MACMG demonstrates robust performance improvements, additional validation on large-scale real-world LMS ecosystems and longitudinal semester-level deployments remains necessary to further assess generalization robustness. Despite impressive performance, the current framework remains limited by the availability of large-scale multimodal educational datasets and requires further validation in real-world semester-long deployments.

VII. DISCUSSION AND FUTURE WORK

A. Discussion of Experimental Findings

The experimental results provide strong evidence that adaptive multimodal personalization is particularly beneficial under educational cold-start conditions. Across both StudentLife and DAiSEE, MACMG consistently outperformed static multimodal Transformers, conventional fine-tuning strategies, and full-model meta-learning approaches. These findings confirm that modality relevance varies substantially across learners and that fixed fusion policies are insufficient for capturing individualized educational risk patterns.

The superior performance of MACMG over StaticMMT demonstrates the limitations of globally optimized multimodal fusion strategies, which assume homogeneous modality importance across all students. By contrast, the proposed meta-adaptive gating mechanism dynamically adjusts modality weights according to learner-specific interaction characteristics, resulting in more discriminative learner representations and improved risk detection accuracy.

The comparison with MAML-Full is particularly informative. Although both approaches leverage meta-learning principles, MACMG adapts only the lightweight gating network while preserving the lower Transformer layers. This selective adaptation strategy reduces the risk of overfitting and enables more sample-efficient personalization under severe data sparsity. The observed performance gains support previous findings by Finn et al. (2017) regarding the effectiveness of meta-learning for rapid adaptation while extending these benefits to multimodal educational environments.

The intervention selection results further highlight the practical significance of adaptive modality weighting. Unlike previous educational prediction models that focus exclusively on learner-state estimation, MACMG directly links multimodal representations to intervention-aware decision support. This capability represents an important step toward autonomous educational digital twins capable of not only predicting risk but also supporting personalized pedagogical action.

Overall, the findings demonstrate that combining multimodal Transformers, educational digital twins, adaptive gating, and few-shot meta-learning within a unified architecture provides a promising pathway toward scalable and personalized educational intelligence systems.

Although MACMG demonstrated strong robustness down to Nsupp = 5 support windows, performance under extreme near-zero-shot conditions (Nsupp = 1–2) remains an open research question and will be investigated in future work.

B. Ethical Implications, Privacy, and Algorithmic Fairness in Biometric Monitoring

Central to the contributions of this work is the deployment of physiological modalities, such as facial expression analysis, which inherently raise significant ethical and privacy concerns. The real-time capture and processing of biometric data, even within controlled educational environments, presents risks related to data security, student surveillance, and potential misuse of affective information [18]. Our proposed MACMG framework, by explicitly modulating the reliance on this sensitive modality through adaptive gating, may inadvertently introduce a mechanism for privacy-preserving operation.

Nevertheless, the permanent storage and processing paradigms of this data stream warrant careful consideration. In future work, we plan to integrate formal privacy guarantees into the MACMG framework. One promising direction is the incorporation of differential privacy constraints directly into the bilevel meta-learning objective [19]. By perturbing the gradients

during the outer-loop meta-update, we could train the gating network initialization to be robust to privacy leakage, ensuring that the adapted student-specific gating weights do not inadvertently encode identifiable information about the historical training population.

Another potential direction involves designing the meta-prior vector $\mathbf{p}_t$ to be ephemeral. Rather than storing a persistent student-specific prior for long durations, the system could periodically reset the meta-prior to a population-level baseline and then adapt it over a limited number of subsequent windows. This would limit the temporal window of biometric data that could be associated with a specific student, enhancing privacy. Moreover, the fairness of the adaptive gating policy across different demographic and socioeconomic groups must be rigorously evaluated. If the gating network learns, from the population meta-training, to associate certain modalities more strongly with risk for subgroups (e.g., correlating video-based analysis with risk for students from a specific cultural background), it could lead to biased intervention allocation. Future work will incorporate fairness constraints into the meta-objective, penalizing scenarios where the gating weights display statistically significant discrepancies in modality reliance between predefined student groups [20].

C. Computational Scalability and Latency Challenges in Real-Time Online Adaptation

While the adaptation time of MACMG (0.47 seconds per new student) is orders of magnitude faster than full-model approaches, the computational demands for deployment at scale in a learning management system (LMS) serving thousands of concurrent students remain a significant engineering challenge. The continuous meta-prior update procedure described in Section IV-D, which performs a few gradient steps every $T_{\text{update}} = 50$ windows impose a recurring computational load. For an LMS with 10,000 active students, this would require processing up to 200 incremental adaptation updates per second, which could strain centralized GPU resources.

Several strategies can be explored to mitigate these latency bottlenecks. First, we can investigate the feasibility of a lazy adaptation protocol where meta-prior updates are skipped for students whose risk score remains consistently low across multiple intervals. This would concentrate computational resources on the minority of high-risk students whose gating policies require rapid refinement. Second, we can explore the application of knowledge distillation techniques [21] to create a smaller, student-specific gating network distilled from the meta-learned initialization. This compressed student sub-model could be deployed on edge devices (e.g., student laptops) to perform local, privacy-preserving adaptation without constant server communication. A third direction involves examining the trade-offs between the frequency of meta-prior updates (T_{update} and W) and adaptation quality. Our current configuration ($T_{\text{update}} = 50, W = 20, K_{\text{inf}} = 3$) was empirically chosen; future work could formalize this as a dynamic programming problem where the system allocates a limited computational budget per student over time to maximize the marginal benefit of each meta-prior update in reducing the uncertainty of the risk prediction.

Furthermore, the computational savings from our first-order MAML approximation, while substantial, could potentially be enhanced by leveraging modern hardware accelerators for matrix operations. The compact nature of the gating network MLP makes it an ideal candidate for deployment on tensor processing units (TPUs) or field-programmable gate arrays (FPGAs) designed for low-latency inference. Exploring these hardware-software co-design strategies could reduce the adaptation time from hundreds of milliseconds to microseconds, enabling truly real-time personalization.

D. Cross-Domain Generalization Limits and Longitudinal Stability of Student Digital Twins

A fundamental limitation of our current evaluation is its confinement to two educational datasets (StudentLife and DAiSEE) with pre-defined modalities and annotation schemes. The generalizability of the meta-learned gating policy to entirely new contexts, such as vocational training platforms, medical simulation environments, or adult learning systems, remains an open question. The modality relevance patterns learned by the meta-initialization from a population of university students may not transfer effectively to a scenario where the primary modality is, for instance, haptic feedback signals from a simulation controller. This challenge highlights a key limitation of in-domain meta-learning.

To improve cross-domain generalization, future work could explore a domain-agnostic meta-prior space. Rather than learning a single meta-initialization ϕ_{meta} for all historical tasks, we could learn a set of meta-initializations $\Phi = \{\phi_{\text{meta}}^{(1)}, \ldots, \phi_{\text{meta}}^{(V)}\}$, each representing a 'prototypical' modality relevance archetype. For a new student, the inference could first perform a rapid, zero-shot assignment by clustering the student's initial interaction patterns to the closest archetype, and then run the inner-loop adaptation from that archetype's initialization. This would allow the system to leverage prior knowledge from diverse learning environments without being confined to a single population's distribution. This approach aligns with recent work on meta-learning for multi-tasking [22, 27, 29].

Equally important is the question of longitudinal stability of the digital twin over long academic terms (e.g., a full semester of 16 weeks). Our continuous meta-prior update mechanism enables tracking evolving student profiles, but it may accumulate drift over many iterations. If the student's dominant predictive modality gradually shifts from text to video as the semester progresses, the frequent gradient updates from the sliding window should, in theory, allow the gating policy to track this change. However, a scrubbing of the meta-prior using stale data points from weeks ago could contaminate the current representation. Investigating methods for graceful forgetting, such as employing a temporally weighted sliding window that exponentially decays the influence of older interaction windows on the gradient update, is a critical direction for ensuring the long-term fidelity of the student digital twin. We plan to evaluate MACMG over a simulated multi-month interaction scenario with artificially introduced concept drift (a sudden change in the relevance of a modality due to a course phase change, e.g., moving from lectures to group projects) to stress-test its robustness [23, 30].

Finally, the current framework's intervention selection is based on the maximum gating weight. While this provides a direct and interpretable link, it may be overly simplistic. A more sophisticated policy could use the full vector of gating weights as input to a reinforcement learning agent that learns optimal intervention strategies through trial and error, using the digital twin's risk score as a reward signal. Extending MACMG to serve as a state representation for a reinforcement learning-based intervention scheduler is a natural and promising next step [24, 30].

Although the current framework focuses on adaptive multimodal fusion and few-shot personalization, future extensions should investigate causal educational reasoning, reinforcement-learning-based intervention optimization, and large language model integration for semantic cognitive state modeling. Furthermore, while the adaptive gating policy captures strong associations between modality relevance and intervention selection, the current framework does not establish formal causal relationships between intervention actions and long-term academic outcomes.

VIII. Conclusion

This study introduced the Meta-Adaptive Cross-Modal Gating (MACMG) mechanism, a novel framework designed to address the critical cold-start problem in multimodal digital twin systems for early detection of at-risk students. By replacing static cross-modal attention layers with a lightweight, student-specific gating network whose initialization is meta-learned, we demonstrated that personalized fusion policies can be efficiently adapted from as few as five initial interaction windows. The bilevel meta-learning objective, built upon a first-order MAML approximation, enables the gating network to capture population-level inductive biases about modality relevance while remaining computationally amenable to rapid per-student personalization. Through extensive evaluation on two multimodal educational datasets, MACMG achieved statistically significant improvements over both static multimodal transformers and full-model meta-learning approaches, with AUROC gains of up to 10 percentage points under extreme data sparsity. Critically, the framework's ability to adapt in under half a second makes it viable for real-time deployment in learning management systems, while the direct coupling between gating weights and intervention selection ensures that personalized pedagogical actions are grounded in the student's most predictive communicative channel. Beyond its empirical performance, MACMG establishes a principled foundation for few-shot cross-modal fusion in educational digital twins, demonstrating that meta-learned parameterized gating offers a data-efficient and computationally practical pathway toward individualized, early intervention. More broadly, this work contributes toward a new generation of autonomous educational digital twins capable of adaptive reasoning, individualized intervention, and continuous learning trajectory optimization in AI-augmented educational ecosystems. Despite these promising results, several limitations remain. The evaluation was restricted to two educational datasets and partially relied on synthetic physiological embeddings to simulate multimodal educational environments. Furthermore, intervention routing was assessed using proxy intervention labels rather than longitudinally validated educational outcomes. Although the results suggest strong personalization capabilities, additional cross-institutional validation is required to fully assess generalization across diverse educational settings. These limitations provide important directions for future research. Future work will investigate the integration of large language models for semantic cognitive-state reasoning and conversational educational orchestration.

References

[1] J. Li, G. Wang, and C. Wang, "Transformer-based sequential modeling framework for personalized learning trajectory analysis," in Proc. Int. Conf. Artificial Intelligence and Education, 2025.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, p. 39, 2019.

[3] A. Sabuncuoglu and T. M. Sezgin, "Developing a multimodal classroom engagement analysis dashboard for higher education," Proc. ACM Hum.-Comput. Interact., vol. 7, no. CSCW2, pp. 1–28, 2023.

[4] Y. Wang, "Affective state analysis during online learning based on learning behavior data," Technol. Knowl. Learn., vol. 28, no. 2, pp. 611–628, 2023.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), Florence, Italy, 2019, pp. 6558–6569.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.

[8] M. She, M. Xiao, and Y. Zhao, "Technological implication of the digital twin approach on the intelligent education system," Int. J. Humanoid Robotics, vol. 20, no. 3, pp. 2350015, 2023.

[9] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the digital twin: A systematic literature review," CIRP J. Manuf. Sci. Technol., vol. 29, pp. 36–52, 2020.

[10] M. Furini, V. Freschi, A. M. Vegni, and M. C. Vuran, "Digital twins and artificial intelligence as pillars of personalized learning models," Commun. ACM, vol. 68, no. 2, pp. 52–60, 2025.

[11] A. Saluja, N. Baig, A. Grover, and R. Adlakha, "Leveraging digital learning environments and predictive analytics to develop adaptive early warning systems for at-risk students in higher education," in Transforming Operational Efficiency and Strategic Decision Making in Higher Education. Hershey, PA, USA: IGI Global, 2026, pp. 210–228.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnnostic meta-learning for fast adaptation of deep networks," in Proc. Int. Conf. Machine Learning (ICML), Sydney, Australia, 2017, pp. 1126–1135.

[13] A. G. Khoee, Y. Yu, and R. Feldt, "Domain generalization through meta-learning: A survey," Artif. Intell. Rev., vol. 57, no. 1, pp. 1–39, 2024.

[14] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5149–5169, 2022.

[15] M. Abdollahzadeh, T. Malekzadeh, and N.-M. Cheung, "Revisit multimodal meta-learning through the lens of multi-task learning," arXiv preprint arXiv:2110.14202, 2021.

[16] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, p. 102274, 2023.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.

[18] Y.-Z. Lin, A. H. J. Alhamadah, M. W. Redondo, K. H. Patel, et al., "Transforming engineering education using generative AI and digital twin technologies," arXiv preprint arXiv:2411.14433, 2024.

[19] M. Á. Rodríguez-Ortiz, J. M. Dodero, and M. S. Ibarra-Sáiz, "Machine learning and generative AI in learning analytics in higher education: A systematic review," Appl. Sci., vol. 15, no. 15, p. 8679, 2025.

[20] G.-G. Lee, L. Shi, E. Latif, Y. Gao, et al., "Multimodality of AI for education: Towards artificial general intelligence," arXiv preprint arXiv:2312.06037, 2023.

[21] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," arXiv preprint arXiv:1701.06538, 2017.

[22] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in Proc. Int. Conf. Machine Learning (ICML), Stockholm, Sweden, 2018, pp. 1568–1577.

[23] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," arXiv preprint arXiv:1803.02999, 2018.

[24] S. Doroudi, E. Brunskill, and M. Brunskill, "Where's the reward? A review of reinforcement learning for educational technologies," Int. J. Artif. Intell. Educ., vol. 29, no. 4, pp. 568–620, 2019.

[25] N. C. Kriegeskorte and P. K. Douglas, "Interpreting encoding and decoding models," Curr. Opin. Neurobiol., vol. 55, pp. 167–179, 2019.

[26] H. ALLIOUI and Y. Mourdi, "Artificial intelligence and education: A systematic literature review," Expert Syst. Appl., vol. 233, p. 120500, 2023.

[27] H. ALLIOUI, Y. Mourdi, and I. Oumaira, "Deep learning and educational data mining for student performance prediction: A comprehensive review," IEEE Access, vol. 12, pp. 45211–45239, 2024.

[28] H. ALLIOUI and Y. Mourdi, "Digital twins in intelligent educational ecosystems: Opportunities, challenges, and future directions," Sensors, vol. 24, no. 6, p. 1987, 2024.

[29] Y. Mourdi, H. ALLIOUI, and H. Kaa, "Generative artificial intelligence for adaptive learning systems: Emerging perspectives in higher education," Multimedia Tools Appl., vol. 84, no. 4, pp. 10231–10258, 2025.

[30] H. ALLIOUI, Y. Mourdi, and H. Kaa, "Multimodal educational intelligence systems using transformer architectures and explainable AI," IEEE Access, vol. 13, pp. 55221–55247, 2025.

An Adaptive Smart Business Intelligence Model Based on Enhanced HGO Discovery for Real-Time in Inpatient Care

Hengki[1]*, Rahmat Gernowo[2], Oky Dwi Nurhayati[3]
Doctoral Program of Information Systems-School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia[1, 2]
Business Digital Study Program, Institute of Science and Business Atma Luhur, Pangkalpinang, Indonesia[1]
Department of Computer Engineering, Diponegoro University, Semarang, Indonesia[3]

***Abstract*—The complexity and dynamics of inpatient care require advanced decision support systems that are fast, adaptive, and capable of real-time execution. This study proposes a novel Smart Business Intelligence (SBI) model developed through an enhanced Hierarchy Governance Outlook (HGO) Discovery approach to achieve high-precision inpatient service prioritization. Unlike conventional frameworks, the proposed model combines an adaptive weighting mechanism, a sustainability optimization process, and real-time data integration from heterogeneous clinical and operational sources. The enhanced HGO Discovery model allows for dynamic adjustment of decision parameters in response to evolving hospital conditions, thereby maximizing patient care. The results demonstrate that the proposed model consistently outperforms baseline approaches in predictive accuracy, stability, and computational efficiency. These findings highlight the potential of the proposed framework to support data-driven decision-making and enhance the quality and efficiency of inpatient care delivery in modern healthcare environments. Experimental results show that the proposed model provides more accurate patient prioritization, faster identification of critical conditions, and improved operational efficiency compared to conventional approaches. Therefore, the proposed system offers an effective intelligent healthcare solution to support real-time inpatient care and data-driven medical decision-making in modern hospitals.**

***Keywords*—Smart business intelligence; decision support systems; HGO discovery; inpatient care; hospital**

I. Introduction

Advances in smart business intelligence and decision support systems have significantly driven the adoption of advanced data analytics and intelligent optimization techniques in the sustainable healthcare domain. The combination of heterogeneous data, governance, and big data facilitates more informed, rapid, and data-driven decision-making in real-time, particularly in the highly complex and fast-paced hospital environment [1]. However, existing implementations are largely conventional and often lack the ability to effectively respond to the next steps and the increasing demand for timely and accurate decision-making [2].

Smart Business Intelligence in hospitals represents a complex ecosystem where clinical services, operational processes, and information systems continuously interact. As core entities within these systems, hospitals face challenges in optimizing limited resources, managing performance, and maintaining high standards of patient safety and comfort in care [3]. Despite these demands, many current systems lack integration and tend to produce fragmented information, limiting the ability to achieve comprehensive, evidence-based decision-making [4].

Within this framework, Business Intelligence (BI) should be viewed not simply as a set of analytical tools, but as a strategic paradigm to support complex decision-making processes, particularly in patient care management [5]. Although previous studies have highlighted the effectiveness of BI in improving the efficiency and quality of healthcare services, critical research gaps remain in the development of models that can seamlessly integrate heterogeneous data in real time, dynamically adapt to changing clinical conditions, and optimize patient prioritization mechanisms [6]. Consequently, there is an urgent need for new approaches that combine adaptive modeling, dynamic weighting strategies, and continuous optimization techniques to enable more accurate, responsive, and context-aware decision support systems [7].

In inpatient care, the accuracy and complexity of decision-making are crucial factors because they directly impact the quality of care, patient safety, the effectiveness of medical procedures, and the efficiency of hospital operations [8]. Delays in detecting changes in a patient's condition can increase the risk of complications and death, prolong the length of stay, and even potentially increase mortality rates. Therefore, a technology-based approach is needed that not only provides historical information but also offers real-time predictive and adaptive analytics capabilities to support more precise decision-making [9].

A technological innovation phenomenon experiencing rapid development in supporting the digitalization of healthcare services is Smart Business Intelligence (SBI) [10]. SBI is believed to implement strategic operations that integrate data sets, analytical techniques, and data dashboard visualizations to transform systematic data processing into valuable, precise information that can be used as a basis for data-driven decision-making. Through the implementation of SBI, technology-based healthcare services can gain more comprehensive insights into operational conditions and clinical services in a systematic and integrated manner, thereby improving the effectiveness of overall healthcare management [11].

*Corresponding author.

Based on the problems described, this study proposes an Adaptive Smart Business Intelligence model based on Enhanced HGO Discovery for Real-Time Precision in Inpatient Care. The designed model focuses on the integration of clinical data and hospital operational data into a Smart Business Intelligence system that has the capabilities of real-time data processing, adaptive analysis, patient condition prediction, and the provision of strategic information precisely and accurately [12]. The implementation of this approach is expected to support medical personnel and hospital management in improving the quality of decision-making, accelerating the handling of changing patient conditions, increasing the efficiency of health services, and reducing mortality rates [13].

II. Related Works

The development of Smart Business Intelligence (SBI) and the latest data-driven health analytics has had a significant impact on the transformation of the modern healthcare system towards a smarter, more integrated, and responsive environment [14]. Recent studies have shown that the integration of Artificial Intelligence (AI) technology with SBI has great potential to increase hospital operational efficiency, improve the quality of patient care, and support clinical decision-making processes more accurately, quickly, and in real-time [15]. However, previous studies still face several challenges, particularly in integrating heterogeneous data, improving the system's ability to adapt to dynamic changes in patient conditions, and implementing intelligent optimization methods that can work adaptively in the complex and ever-evolving inpatient care environment [16].

The role of Smart Business Intelligence in the transformation of modern healthcare services is increasingly significant with the integration of artificial intelligence through the use of predictive analytics and intelligent clinical decision support systems [17]. Numerous research studies have shown that the application of artificial intelligence can accelerate diagnosis, analyze patient risks, and improve the overall effectiveness of patient care. Therefore, most existing research focuses on general AI implementation and has not yet integrated the concept of Smart Business Intelligence, supported by adaptive optimization mechanisms, into a real-time inpatient care environment [18].

Developing a real-time patient monitoring system based on Artificial Intelligence (AI) and the Internet of Things (IoT) to support continuous patient condition monitoring [19]. The designed system is capable of integrating primary health data with clinical data to detect changes in patient conditions more quickly and accurately. The results of this study demonstrate the effectiveness of the patient monitoring process, but the research still has limitations in supporting strategic decision-making based on Smart Business Intelligence (SBI) and has not implemented an intelligent optimization mechanism to determine adaptive patient care priorities [20].

The Smart Hospital Framework, which integrates Business Intelligence technology into hospital analytics, was developed to improve operational efficiency and the quality of healthcare services for inpatients [21]. Research results show that the use of a BI-based analytical dashboard can help hospital management monitor service performance and resource utilization more effectively and in a structured manner. However, the applied approach tends to be static, thus not being able to support adaptive analysis or predict patient conditions in real time [22].

In the context of developing patient prediction systems in hospital settings, various studies have developed inpatient condition prediction models using deep learning algorithms. The results show that this approach can improve the accuracy of predicting critical patient conditions compared to conventional methods. However, the implemented models still focus on simple analyses and have not yet integrated comprehensive Business Intelligence mechanisms to support clinical and operational decision-making processes in hospitals [23].

Developing a hybrid optimization technique approach to improve the performance of an artificial intelligence-based healthcare analytics system. The study demonstrates that a combination of heuristic algorithms and adaptive optimization can improve the accuracy of healthcare data classification while accelerating the process of finding optimal solutions for complex datasets. Despite making significant contributions to the development of healthcare optimization, the study has yet to implement the optimization method within an Adaptive Smart Business Intelligence framework to support real-time inpatient care [24].

Based on the latest research, it is understood that the application of Business Intelligence (BI), Artificial Intelligence (AI), and optimization techniques in the healthcare sector has had a significant impact on improving the quality of hospital services. The integration of these technologies can support a more efficient, accurate, and data-driven decision-making process. However, there are still a number of research gaps that have not been fully addressed, especially related to the integration of clinical and operational data in real time, the system's ability to adapt to dynamic changes in patient conditions, optimization of service priorities, and the implementation of adaptive heuristic methods in Smart Business Intelligence systems. Therefore, this study proposes an Adaptive Smart Business Intelligence Model Based on Enhanced HGO Discovery for Real-Time Precision in Inpatient Care as an innovative approach that combines adaptive analytics, real-time heuristic optimization, and intelligent prediction mechanisms to support more precise, responsive, and efficient inpatient care.

III. Research Methodology

This research object uses a quantitative approach based on system development (system development approach) to design an Adaptive Smart Business Intelligence Model Based on Enhanced HGO Discovery for Real-Time Inpatient Care. The research process begins with the collection of clinical data and hospital operational data sourced from Electronic Health Records (EHR) and a real-time patient condition monitoring system where the total patient data used was 2,780 samples. Next, the data processing stage is carried out which includes three main components, namely Hierarchy (H), Governance (G), and Outlook (O), to support the process of data integration, information management, and adaptive analysis in improving the effectiveness of inpatient services. The implementation stages of the Enhanced HGO Discovery Framework in this study are carried out through several systematic processes in Fig. 1.

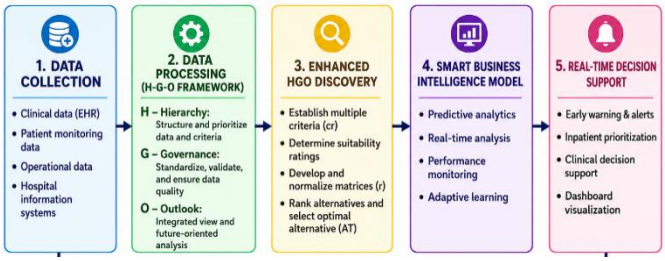


Fig. 1. Methodology framework.

The first stage involves defining multiple decision criteria (criteria/cr) that serve as the primary basis for evaluating inpatient service conditions and supporting clinical decision-making. These criteria are designed to represent important aspects of patient conditions, hospital operations, and healthcare service priorities within the Smart Business Intelligence framework.

The second stage focuses on determining the suitability rating for each alternative based on the established criteria. In this process, the evaluation values are represented through the components of Hierarchy (H) and Governance (G), which are used to classify data structures, prioritize patient conditions, and manage healthcare information systematically. This stage enables the model to identify the relevance and importance of each alternative in supporting adaptive decision-making.

The next stage involves constructing several decision matrices derived from the defined criteria. These matrices are subsequently processed using a normalization procedure based on positive and negative attribute consensus values to produce balanced and standardized data representations. The normalization process generates normalized matrix values symbolized as (r), which are utilized to improve analytical consistency, reduce data bias, and enhance computational efficiency in the optimization process.

Finally, the ranking and selection process is performed by calculating the weighted aggregation of the normalized matrix values. Each alternative is evaluated based on its calculated risk and performance score, where the alternative with the lowest risk value and highest optimization performance is selected as the optimal alternative (AT). This process enables the proposed model to support real-time adaptive decision-making in inpatient care through accurate prioritization and intelligent analytical recommendations.

IV. Result and Analysis

This study develops an Adaptive Smart Business Intelligence model based on Enhanced HGO Discovery, designed to support real-time decision-making in inpatient care services in hospitals. The proposed model integrates various hospital data sources, namely Electronic Health Records (EHR), patient monitoring data, and hospital operational data, all combined into a single adaptive Business Intelligence platform. Implementation results show that the system is capable of processing data quickly and continuously, thus supporting a more effective and accurate decision-making process in a dynamic healthcare environment.

The research phase begins with the Data Collection process, which serves as the primary foundation for developing a Smart Business Intelligence system. In this phase, various types of data are collected from several sources that serve as reference points for the hospital. The collected data includes Electronic Health Records (EHR), operational data, and information originating from the hospital. Information in the EHR includes patient health histories, medical diagnoses, treatment procedures, and medication administration history. Furthermore, this study utilizes not only clinical data but also hospital operational data, such as bed availability, patient emergencies, healthcare provider schedules, and patient queue data to support a more comprehensive and accurate decision-making process. We look at five key processes: (A) Data Collection, (B) Data Processing HGO Discovery, (C) Enhanced HGO Discovery, (D) Smart Business Intelligence Model, (E) Real Time Decision Support.

A. Data Collection

The initial stage of the research began with the Data Collection process, which serves as the primary foundation for developing an Adaptive Smart Business Intelligence system. This stage involved data collection from hospitals to generate comprehensive and accurate information. The data collected included Electronic Health Records (EHR). This study utilized 2,780 datasets, consisting of inpatient data, patient condition monitoring data, and hospital operational data collected during the study period. The characteristics of the dataset can be seen in Table I.

TABLE I. Characteristics of the Dataset

Specifications	Description
Total number of datasets	2,780
Time slot	2023-2026
Attribute of positive value	Insurance Provider, Room Class, Admission Type, Severity Score, Test Result
Attribute of negative value	Surgery

The EHR datasets utilized in this study contained a wide range of essential information regarding patient conditions, including medical histories, laboratory examination results, clinical diagnoses, treatment procedures, and medication administration records. In addition, the system collected real-time data directly from patient monitoring devices. Beyond clinical information, the research also incorporated hospital operational data, such as inpatient room capacity, bed occupancy availability, and patient queue information, to facilitate a more comprehensive and effective decision-making process.

B. Data Processing HGO Discovery

HGO Discovery is a multidimensional data processing method that integrates: priority structure (Hierarchy), control and decision rules (Governance), and strategic prediction and insight (Outlook). Table II shows implementation of HGO Discovery.

TABLE II. IMPLEMENTATION PROCESSING HGO DISCOVERY

Model	Category	Interpretation
Priority structure (Hierarchy)	Level 1	Main objective
	Level 2	Decision criteria
	Level 3	Sub-criteria
	Level 4	Decision alternatives
Control and decision rules (Governance)	Data validation	Rule-based filtering
Strategic prediction and insight (Outlook)	Prediktif	Intelligence forecasting

After an agreement is reached regarding the weight of each attribute or criterion determined by the experts, the next stage is to assign CRISP values to the data based on the previously determined hierarchy. There are six criteria (Cr) used in the assessment or decision-making process. The first criterion, Cr1 (Insurance Provider), is a benefit criterion with a weight of 0.10, meaning the better the insurance provider, the higher the score. The second criterion, Cr2 (Surgery), is a cost criterion with a weight of 0.20, so a lower score is considered better. Next, Cr3 (Room Type) has a benefit characteristic with a weight of 0.175, where a better room type gives a higher score. Cr4 (Admission Type) is also a benefit criterion with a weight of 0.125. Then, Cr5 (Severity Score) is a benefit criterion with the highest weight along with several other criteria, namely 0.20, indicating a high level of importance in the assessment. Finally, Cr6 (Test Result) is a benefit criterion with a weight of 0.20, so a better test result will increase the final score. All of these criteria's weights are used to determine the priority or overall evaluation results based on their respective levels of importance. Fig. 2 shows specification and value attributes of the characteristic dataset.

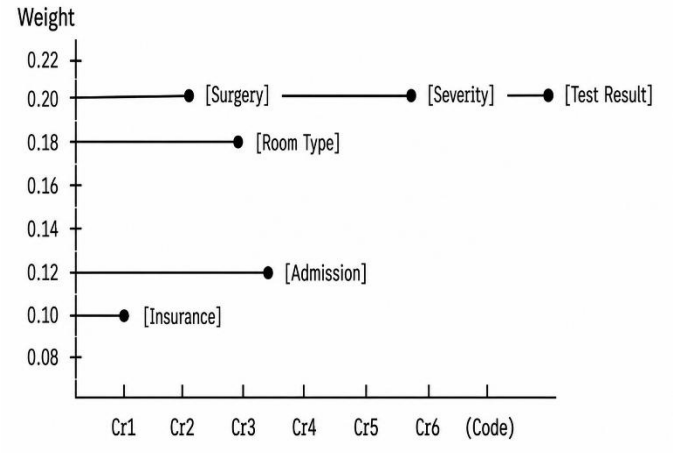


Fig. 2. Specification and value attribute.

C. Enhanced HGO Discovery

Enhanced HGO Discovery demonstrates superior performance in finding optimal solutions compared to conventional methods. Improvements to the HGO algorithm are made through the addition of optimization mechanisms and parameter adjustments to make the exploration and solution exploitation processes more efficient. Based on testing results, this method is able to provide more precise, stable, and consistent preference values in selecting the best alternative based on the criteria of Insurance Provider, Surgery, Room Type, Admission Type, Severity Score, and Test Results.

This stage also shows a faster convergence rate in the calculation process, resulting in more efficient computing time. By considering the weight and type of criteria (benefit and cost), the Enhanced HGO Discovery model successfully provides more objective evaluation results and supports more accurate decision-making, Fig. 3 shows HGO Discovery process.

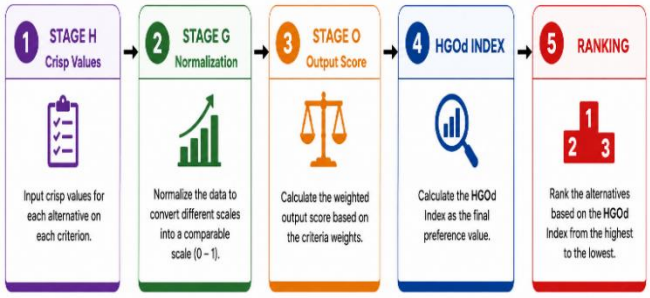


Fig. 3. HGO discovery processes.

Therefore, this model is considered effective for application in decision support systems, especially in the process of analyzing and selecting the best alternative in the healthcare sector. Table III shows the stage enhanced HGO Discovery.

TABLE III. ENHANCED HGO DISCOVERY METHOD

Stage	Description	Formula
Crisp Values	Input crisp values for each alternative on each criteria	Cr = criteria (Cr1 to Cr..n)
Normalization	Normalize the data to convert different scale into a comparable scare (0-1)	Value scale between 0 to 1
Output Score	Calculate the weighted output score based on the criteria weights.	Total weight =1.00
HGO Discovery Index	Calculate the HGO Discovery index as the final preferenced value	HGO Discovery index or HGOD Index, the highest priority value is 0.000, and the lowest priority value is 1
Ranking	Rank the alternative based on the HGO Discovery index from the higest to the lowest	Alternative are ranked from the highest to the lowest HGOD index.

The crisp data representing the criteria values contains information about each criterion and its weight. These crisp values are optional and serve as a reference limit for determining the value range for each criterion. Each crisp value has a specific weight derived from the conversion process, which is then applied to the HGO Discovery model calculations. The weighting process also influences the nature of the criteria, which are divided into two categories: positive and negative factors. Meanwhile, alternative values are used to store the evaluation results of each alternative based on all available criteria data.

Normalization attributes must be carried out so that the important process in compiling data in the stage model that occurs in HGO Discovery. Next stage is output score, where the weighted output score based on the criteria weights is calculated.

The output and HGOD Index have different scales, but each serves a different purpose in the decision support framework of this model, where the higher value output is the best value, while with the HGOD index, the lower value is better. Ranking is based on the HGOD index, where the lowest value = rank 1. This is useful for prioritizing patients based on the HGO discovery metric. Table IV shows different output score and HGOD Index.

TABLE IV. DIFFERENT OUTPUT SCORE VS HGOD INDEX

Interpretation Output	Value
Output Score (Simple Additive Weighting)	Higher output score indicates better patient condition overall. Range: 0–1. Calculated via weighted sum of normalized criteria.
HGOD Index	Lower HGOd index = higher inpatient priority. Formula: 1 / Σ(Wj × Xj). Patients in top-left scatter quadrant need immediate attention.

D. Smart Business Intelligence Model

After patient data is collected, processed using the HGO Discovery, and analyzed using the Enhanced HGO Discovery method, the system then enters the Smart Business Intelligence stage to generate more adaptive and strategic insights.

At this stage, the system utilizes previous calculation results, such as normalization values, output scores, and the HGOd Index, to perform predictive analytics. This analysis is used to estimate patient conditions, risk levels, service priorities, and potential changes in a patient's health status during the inpatient process. With this approach, hospitals can take faster and more accurate action based on data automatically analyzed by the system.

This Smart Business Intelligence Model applies the concept of adaptive learning, which is the system's ability to adjust its analysis process based on the latest data patterns. The system can learn from previous patient data, making recommendations and prioritization more accurate over time. This adaptive approach makes the model more flexible in responding to changing patient conditions and hospital operational dynamics. The Smart Business Intelligence Model serves as the center of intelligent analysis in this study because it integrates data, Enhanced HGO Discovery, and real-time analytics to produce faster, more accurate, and more efficient decisions in inpatient care services.

The system utilizes previous calculation results, such as normalization values, output scores, and the HGOd Index, to perform predictive analytics. This analysis is used to estimate patient conditions, risk levels, service priorities, and potential changes in a patient's health status during the inpatient process. With this approach, hospitals can take faster and more accurate action based on data automatically analyzed by the system. Fig. 4 shows Smart Business Intelligence Output Score and HGOD Index.

CODE	NAME	AGE	GENDER	OUTPUT	HGOD INDEX	RANK	PRIORITY
P000011	Patient A	24	Male	0.6225	0.013841	326	Critical
P000010	Patient B	89	Female	0.6575	0.017937	2201	Medium
P000009	Patient C	45	Female	0.6850	0.017094	1808	Medium
P000008	Patient D	52	Male	0.5925	0.014440	549	Critical
P000007	Patient E	63	Male	0.5125	0.016327	1487	High
P000006	Patient F	55	Male	0.6850	0.012739	86	Critical
P000005	Patient G	66	Male	0.5625	0.021622	3025	Low
P000004	Patient H	66	Male	0.7000	0.016667	1626	High
P000003	Patient I	18	Male	0.6575	0.017937	2200	Medium

Fig. 4. Smart business intelligence output score and HGOD index.

E. Real Time Decision Support

Real-Time Decision Support is the final stage in the Adaptive Smart Business Intelligence model based on Enhanced HGO Discovery, which functions as a direct and automated decision support system in inpatient care. At this stage, all analysis results from the previous process—from data processing, Enhanced HGO Discovery calculations, to the Smart Business Intelligence Model—are used to generate fast, accurate, and adaptive decision recommendations based on real-time patient conditions.

Inpatient prioritization is used to determine the priority of inpatient services based on the HGOd Index and patient data analysis. Patients with higher risk levels receive priority treatment over lower-risk patients. This mechanism helps hospitals manage service queues, treatment room utilization, and medical resource allocation more efficiently and objectively.

This stage also supports clinical decision support, which provides data-driven recommendations to assist doctors and healthcare professionals in making medical decisions. The system can provide insights into patient severity, health condition analysis results, and prioritized action recommendations based on automatically processed data. This way, medical decisions no longer rely solely on manual observation but are also strengthened by intelligent data analysis.

Another feature is the visualization dashboard, an interactive visual display that presents patient information, priority scores, risk status, and analysis results in real time. This dashboard makes it easier for hospital management and medical personnel to quickly and comprehensively monitor patient conditions. Clear data visualization also contributes to more effective monitoring and evaluation of healthcare services. Fig. 5 show dashboard interactive visual that priority score.

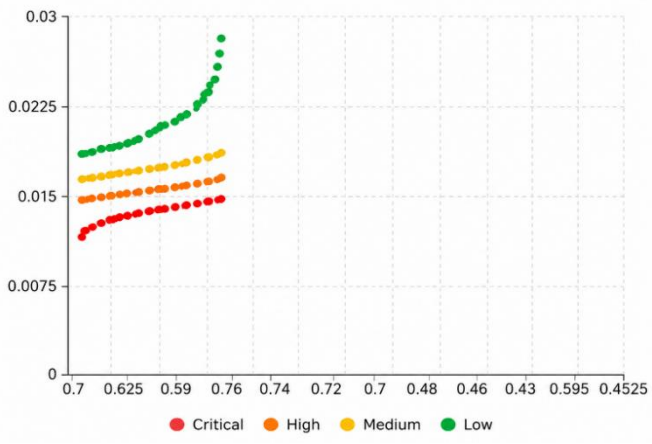


Fig. 5. Dashboard interactive visual for the priority score.

V. Conclusion

An HGO Discovery-based model supports real-time decision-making in inpatient care. The proposed model integrates data collection, the HGO Discovery model, and a Smart Business Intelligence system to generate more adaptive, accurate, and efficient analysis. Implementing this method can improve the patient evaluation process by calculating preference scores and the HGOd Index based on various criteria.

The integration of predictive analytics, real-time analysis, adaptive learning, and real-time decision support features enables the system to provide early warnings, dashboard visualizations, and automated patient service priority recommendations. Enhanced HGO Discovery has also been shown to increase the effectiveness of the solution exploration and exploitation process compared to conventional methods, resulting in more stable and consistent decisions. Thus, this research provides an important contribution to the development of a modern healthcare business intelligence system that can improve hospital service quality, operational efficiency, and rapid response to inpatient conditions dynamically and in real-time.

For further research, it is hoped that the Adaptive Smart Business Intelligence model based on Enhanced HGO Discovery can be developed by adding more variables and more complex health data sources, such as electronic health records (EHRs), health IoT sensor data, and continuous patient monitoring data. Further development can also be done by integrating machine learning and deep learning technologies so that the system can produce more accurate patient condition predictions and adapt to real-time data changes. In addition, further research is expected to test this model on a larger hospital scale and different health environments to determine the level of effectiveness, stability, and flexibility of the system in real conditions.

Acknowledgment

Special recognition is given to the healthcare institutions and organizations that provided support, facilities, and data related to inpatient care. Their collaboration and contributions significantly assisted in the analysis and implementation of the adaptive smart business intelligence model developed in this research.

References

[1] Alowais, S., et al. “Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice.” *BMC Medical Education*, 23(1), 689, 2023.

[2] Nguyen, H., et al. “Real-Time Patient Monitoring Using AI-Driven Healthcare Analytics.” *Sensors*, 23(14), 6211, 2023.

[3] Singh, R., et al. “Smart Hospital Framework Integrating Business Intelligence and AI Systems.” *IEEE Access*, 11, 112345–112360, 2023.

[4] Patel, V., et al. “Deep Learning-Based Inpatient Deterioration Prediction Models.” *Journal of Medical Systems*, 47(5), 77, 2023.

[5] Rahman, M., et al. “Hybrid Optimization Techniques for Intelligent Healthcare Analytics.” *Applied Soft Computing*, 156, 111450, 2024.

[6] Luo, H., et al. “Real-Time AI Frameworks for Hospital Operational Analytics.” *IEEE Journal of Biomedical and Health Informatics*, 28(2), 1033–1045, 2024.

[7] Sharma, P., et al. “Smart Healthcare Ecosystems Integrating AI and Business Intelligence.” *Healthcare Analytics*, 5, 100311, 2024.

[8] Chen, X., et al. “Adaptive Business Intelligence Systems for Real-Time Healthcare Decision Support.” *Journal of Healthcare Engineering*, 2025.

[9] Mimmo, S. S. “Engineering Intelligent Health Systems: AI-Powered Business Analytics for Informed Clinical Decision-Making.” Department of Educational Technology and Engineering, Gazipur Digital University, Gazipur, 2025.

[10] Al-Nafjan, A., et al. “Artificial Intelligence in Predictive Healthcare: A Systematic Review.” *Journal of Clinical Medicine*, 2025.

[11] Irene, N. J. “Artificial Intelligence in Predictive Healthcare Analytics.” *IAA Journal of Biological Sciences*, 13(1), 67–74, 2025.

[12] Alghamdi, A., Alfakeeh, A., & Alsubaie, N. “Artificial Intelligence for Smart Healthcare Systems.” *Healthcare*, 12(2), 145, 2024.

[13] Zhang, X., Liu, Y., & Wang, H. “Real-Time Healthcare Analytics for Inpatient Monitoring Using AI Techniques.” *Journal of Biomedical Informatics*, 149, 104567, 2024.

[14] Moreno, E., & Garcia, F. “Adaptive Optimization Models in Smart Healthcare Environments.” *Applied Soft Computing*, 154, 111315, 2024.

[15] Hossain, M. S., Muhammad, G., & Guizani, N. “AI-Driven Healthcare Systems for Smart Hospitals.” *Future Generation Computer Systems*, 154, 381–392, 2024.

[16] Ibrahim, A., Hassan, M., & Ali, S. “Enhanced Metaheuristic Algorithms for Healthcare Optimization Problems.” *Soft Computing*, 29, 4451–4470, 2025.

[17] Li, Q., Sun, W., & Huang, Y. “AI-Powered Smart Inpatient Care Monitoring Framework.” *Healthcare Technology Letters*, 12(1), 44–53, 2025.

[18] Fernandez, R., & Gomez, J. “Real-Time Predictive Analytics in Smart Healthcare Systems.” *Journal of Medical Systems*, 49(3), 55, 2025.

[19] Park, S., Lee, D., & Kim, J. “Hybrid Optimization Algorithms for Intelligent Healthcare Applications.” *Applied Intelligence*, 55, 6678–6695, 2025.

[20] Wu, Y., Zhang, H., & Li, X. “Advanced Decision Support Models for Inpatient Care Analytics.” *Healthcare Analytics*, 2026.

[21] Kumar, R., & Patel, D. “Enhanced Swarm Intelligence Algorithms for Healthcare Optimization.” *Applied Soft Computing*, 2026.

[22] Hassan, R., Ali, M., & Ibrahim, S. *Business Intelligence Frameworks for Smart Hospital Systems*. Journal of Healthcare Engineering, 2026.

[23] Ahmed, T., Rahman, M., & Noor, F. *Next-Generation Adaptive Healthcare Intelligence Systems Using AI and Optimization*. Artificial Intelligence in Medicine, 145, 102781, 2026.

[24] Zhou, T., Li, H., & Xu, Y. *Optimization Algorithms in Smart Healthcare Analytics: A Review*. Artificial Intelligence Review, 57, 88, 2024.

Multi-Relation Knowledge Graph-Guided Transformer with Objective-Aware Route Selection for Vehicle Routing Problems with Time Windows

Somkiat Kosolsombat[1], Chiabwoot Ratanavilisagul[2*]
Data Science and Innovation Program, College of Interdisciplinary Studies, Thammasat University, Thailand[1]
Faculty of Applied Science-Department of Computer and Information Science,
King Mongkut's University of Technology North Bangkok, Thailand[2]

***Abstract*—Vehicle routing problems with time windows require route construction methods that can reason jointly about spatial distance, vehicle capacity, service time, and customer time windows. Recent neural combinatorial optimization studies show that Transformer and graph-based models can learn routing policies, but many models still rely mainly on node features and do not expose which routing relations are responsible for feasibility and route quality. This study proposes a multi-relation knowledge graph guided Transformer framework for the vehicle routing problem with time windows. For each instance, a pairwise knowledge graph is built from spatial, temporal, capacity, slack, and depot identity relations. These relation vectors are projected into head-specific attention biases inside a pointer-style Transformer, while a feasibility mask enforces capacity and time window constraints during decoding. The framework is trained by imitation from OR-Tools teacher routes and evaluated on the 56 Solomon benchmark instances and on synthetic Solomon-style extensions. The original Transformer baseline decoded feasible routes for 54 of 56 Solomon instances, whereas the full knowledge graph guided model decoded feasible routes for all 56 instances. Additional controlled AM style, POMO style multistart, and edge-aware Transformer baselines improve route quality gaps on feasible instances but still leave one Solomon instance infeasible, highlighting the proposed framework's feasibility-oriented role. On three 120 instance synthetic runs, objective-aware route selection over relation-specialized models reached a mean feasible rate of 0.997 and reduced the mean distance gap by 125.418 Solomon coordinate distance units compared with the relation-free Transformer baseline. The contribution is therefore a diagnostic and extensible framework that integrates explicit routing knowledge into neural attention, evaluates relation ablations, and combines specialized route candidates according to the target operational objective.**

***Keywords*—Vehicle routing problem with time windows; knowledge graph; Transformer; neural combinatorial optimization; relation aware attention; route selection**

I. Introduction

The vehicle routing problem with time windows (VRPTW) is a central problem in logistics optimization. A fleet must serve geographically distributed customers while satisfying vehicle capacity and customer service intervals. The Solomon benchmark remains a standard controlled testbed for this setting [1]. Constraint and operations research solvers, including OR-Tools, provide strong references because they represent routing constraints directly [2].

Learning-based solvers are attractive when repeated route construction, fast policy reuse, or learned route priors are required. Recent surveys describe rapid growth in neural combinatorial optimization, reinforcement learning, and learning to route methods [3]-[5]. However, a practical gap remains: many neural VRP models learn from node features without explicitly separating the spatial, temporal, capacity, and slack relations that drive VRPTW feasibility.

In this study, a pointer Transformer refers to an attention-based encoder-decoder route construction model whose output pointer distribution selects the next customer from the current set of feasible nodes. The model, therefore, predicts a sequence of customer visits rather than a fixed class label.

This work studies whether a knowledge graph (KG) can operate as a structured control layer for a Transformer routing model. Instead of treating the KG as a post hoc explanation, the proposed framework injects pairwise relation information directly into Transformer attention. The study is positioned as a framework and diagnostic analysis rather than as a claim that a compact neural model outperforms OR-Tools.

The contributions are: a multi-relation KG representation for VRPTW instances; a KG-guided pointer Transformer that converts relation vectors into attention bias; relation ablations and adaptive variants that reveal how relation groups affect decoded routes; and objective-aware route selection with multi-seed robustness evaluation.

The rest of the study is organized as follows: Section II reviews neural routing and knowledge guided routing studies. Section III states the research gap. Section IV presents the KG-guided framework, relation definitions, attention bias, feasibility mask, and implementation settings. Section V describes the experimental design, Section VI reports the results, Section VII discusses implications, and Section VIII and Section IX summarize limitations and conclusions.

II. Related Work

Hybrid learning and optimization methods have become common in vehicle routing. Deep reinforcement learning has been used for column generation [6], neural construction heuristics have been embedded in large neighborhood search

*Corresponding author

[7], and combinatorial optimization enriched machine learning has been applied to dynamic VRPTW [8]. Multi-objective and multi-agent variants further expand the decision space beyond distance minimization [9]-[12].

Recent time window and electric vehicle studies use reinforcement learning and graph-structured data to model feasible actions, resource use, and operational constraints [13]-[15]. Transformer-based routing models use attention to represent customer interactions, including collaborative Transformer approaches, EFormer edge-based Transformers, and local global feature fusion Transformers [16]-[18]. GASE graph attention with edge fusion and symmetry preserving graph attention also suggest that edge information can improve routing representations [19], [20]. These models motivate relation-aware routing, but they do not isolate VRPTW spatial, temporal, capacity, slack, and identity relations as separate ablation controls inside a head-specific attention bias.

Broader generalization has been explored through equity-aware Transformers, cross-problem learning, foundation-style route models, mixture of experts, large-scale reinforcement learning, and route foundation models [21]-[26]. Knowledge-guided metaheuristics and hybrid reinforcement learning show that explicit domain knowledge can improve search behavior [27], [28]. Other recent routing research includes GCN-assisted search, hyperheuristic and VAE-based approaches, edge-based learning for VRPTW, and backhaul routing [29]-[32]. Closely related Thai VRPTW studies by Kosolsombat and Ratanavilisagul examined modified ant colony optimization and applied deep reinforcement learning for time window routing [33], [34]. The present work differs by using a pairwise KG as a direct attention bias inside a Transformer and by testing relation groups through ablation and route selection.

III. Research Gap and Design Rationale

The literature indicates three limitations that motivate this study. First, many neural VRP models encode customers primarily as node sequences. This is convenient for attention-based models, but VRPTW feasibility is inherently pairwise: whether customer j can follow customer i depends on travel time, service completion, waiting time, residual capacity, and both time windows. A model that only receives node attributes must infer these pairwise constraints indirectly from training examples.

Second, recent neural routing studies often report final route quality but provide less diagnostic evidence about which kinds of routing knowledge help. For a logistics decision maker, it is useful to know whether spatial proximity, temporal feasibility, capacity compatibility, or slack information is responsible for improved decoded routes. A relation ablation design, therefore, gives more insight than a single black box comparison.

Third, no single learned construction policy is expected to dominate all instance classes. Clustered, random, and mixed Solomon-style distributions have different structural properties. A relation setting that is useful for clustered customers may not be best for dispersed customers with tight windows. This motivates a framework in which relation-specialized models are treated as candidate route generators, and route validation is used before objective-aware selection.

IV. Proposed Framework

Fig. 1 summarizes the proposed pipeline. Each VRPTW instance is converted into node features and a pairwise KG. OR-Tools teacher routes provide imitation labels. A pointer Transformer receives node features, while the KG contributes relation-specific attention bias. The decoder applies a hard feasibility mask, and a candidate pool of relation-specialized models is filtered and ranked by the chosen routing objective.

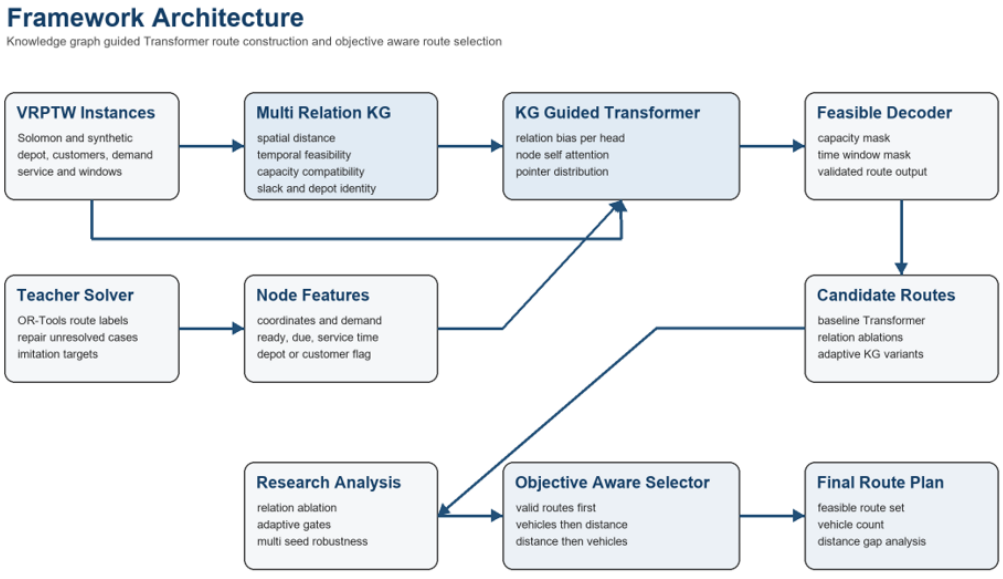


Fig. 1. Proposed multi-relation KG-guided Transformer framework with objective-aware route selection.

A. Problem Setting

Each instance contains one depot and n customers. Customer i has coordinates, demand, service duration, and a time window. A feasible route starts and ends at the depot, respects vehicle capacity, and begins service within every customer's time window. Evaluation reports the feasibility, the number of vehicles, total distance, and the gap relative to the OR-Tools reference.

B. Multi-Relation Knowledge Graph

For each ordered node pair, the KG stores relation features. The implemented groups are spatial relations, temporal feasibility, time window overlap, capacity compatibility, waiting time and slack, and depot or identity indicators. Ablation variants include spatial, temporal, capacity, feasible, slack, no slack, and full relation settings.

Let V={0,...,n} denote the depot and customer nodes, where node 0 is the depot. Customer i has coordinates x_i, demand q_i, service time s_i, ready time e_i, and due time l_i. Vehicle capacity is Q, the depot horizon is H, and d_ij is the Euclidean distance. For every ordered pair (i,j), the KG relation vector r_ij contains normalized distance d_ij/max(d), a near-neighbor indicator 1[d_ij <= q_0.20(d)], temporal feasibility 1[e_i+s_i+d_ij <= l_j], reverse feasibility 1[e_j+s_j+d_ij <= l_i], time-window overlap 1[max(e_i,e_j) <= min(l_i,l_j)], and capacity compatibility 1[q_i+q_j <= Q].

Slack and waiting relations are computed as wait_ij=max(0,e_j-(e_i+s_i+d_ij))/H and slack_ij=clip((l_j-(e_i+s_i+d_ij))/H,-1,1). Identity relations are same_node=1[i=j], source_is_depot=1[i=0], and target_is_depot=1[j=0]. These formulas make the KG deterministic and reproducible from a Solomon-style input file without external semantic resources.

The spatial group represents geometric closeness through normalized distance and near-neighbor indicators. These relations give the model a direct signal about route compactness. The temporal group represents whether one customer can precede another under service and travel time, including reverse feasibility and time window overlap. Slack and waiting relations capture how much scheduling flexibility remains when a customer pair is visited consecutively. The capacity group encodes whether two customer demands can coexist under the vehicle capacity, and the identity group distinguishes depot-related transitions from customer-to-customer transitions.

This formulation is intentionally simple and deterministic. The KG does not require external semantic knowledge or manually curated ontologies. Instead, it transforms the mathematical constraints of VRPTW into pairwise relation channels that can be consumed by the neural model. This makes the framework reproducible on any Solomon-style instance and allows each relation group to be removed or retained in ablation experiments.

C. KG-Guided Pointer Transformer

The relation-free baseline is a pointer-style Transformer that predicts the next customer from node embeddings. The KG-guided version projects relation vectors into head-specific attention terms. These terms bias self attention before the pointer distribution is computed. During decoding, infeasible customers are masked to prevent capacity and time window violations.

Formally, node features X are projected to hidden states h_i. For attention head k, relation vector r_ij is mapped to a scalar bias b_ij^k=tanh(w_k^T r_ij+c_k) gamma, where gamma is a learned scale. The self-attention logit becomes a_ij^k=(Q_i^k K_j^k)/sqrt(d_h)+b_ij^k before the softmax over candidate keys j. In the implementation, the same relation-bias tensor is supplied to each relation-aware encoder layer, allowing different heads to learn different uses of spatial, temporal, capacity, slack, and depot relations.

The attention bias design is important because it lets each attention head learn a different use of the KG. One head may emphasize spatial closeness, another may emphasize temporal feasibility, and another may attend to depot-related transitions. The model, therefore, does not concatenate relation features to node features only once at the input layer; it uses relation information whenever pairwise attention scores are computed.

D. Training and Decoding

Training is formulated as imitation learning from teacher routes. For each partial route state, the teacher provides the next customer in the sequence, and the model is optimized to assign high probability to that teacher action. This design avoids the high variance of pure reinforcement learning and makes the experiment feasible on a modest GPU. It also creates a clear reference point: the neural model is learning to approximate the route construction behavior induced by the OR-Tools teacher.

During inference, the decoder constructs routes greedily under a feasibility mask. The mask removes customers who would violate capacity or time windows from the next action set. When no feasible customer remains, the vehicle returns to the depot, and a new route is started. After decoding, each route plan is validated again before it is used in evaluation or route selection. This validation step is essential because the study measures decoded operational feasibility, not only sequence prediction loss.

The feasibility mask is defined at each partial route state. A customer j can be selected only if j is unvisited, q_j is not larger than remaining capacity, and the service start time max(e_j,t_i+s_i+d_ij) is not later than l_j. The depot action is allowed after at least one customer has been served on the current route, or when no customer satisfies the capacity and time-window tests. This hard mask prevents the decoder from selecting actions that are known to be infeasible before route validation.

E. Objective-Aware Route Selection

Different relation groups can be useful for different instance structures and objectives. Therefore, relation-specialized models generate route candidates. Valid candidates are ranked by either vehicles then distance or distance then vehicles. This route-level step converts relation diversity into an explicit selection mechanism.

Two ranking choices are considered because routing objectives are not always equivalent. A logistics operation that pays a fixed cost per vehicle may prefer vehicles then distance, while an operation with a fixed fleet may prefer distance then vehicles. Reporting both objectives prevents an overly broad claim of superiority and shows how the same candidate pool can be adapted to different operational priorities.

F. Implementation Details

The implementation settings used in the reported experiments are summarized in Table I. The local neural comparison uses the relation-free pointer Transformer as a direct baseline and keeps the same teacher-label generation and decoded feasibility validation as the KG-guided models.

TABLE I. IMPLEMENTATION AND TRAINING SETTINGS

Setting	Reported value
Relation-free model	d_model=128, 8 heads, 4 layers, FFN=256
KG-guided model	d_model=128, 4 heads, 3 relation-aware layers, FFN=256
Regularization	dropout=0.1, weight decay=1e-4, gradient clipping=1.0
Optimizer	AdamW; lr=1e-3 for relation-free baseline, 8e-4 for KG variants
Training	imitation learning from OR-Tools labels; batch size 32; Solomon 80 epochs; synthetic 50 epochs
Decoding	greedy pointer decoding with hard capacity and time-window mask
Controlled neural baselines	AM style attention pointer; POMO style 16-start selection; edge-aware Transformer with full relation features
Hardware	Remote GPU: NVIDIA RTX 3050, CUDA 12.4, PyTorch 2.6.0+cu124

G. Competitive Benchmarking Scope

Table II positions the proposed framework against established neural routing families and adds controlled neural baselines that were run under the same Solomon data, teacher label, decoding, and validation protocol. The AM style, POMO style multistart, and edge-aware Transformer rows are locally controlled baselines rather than official reproductions of AM, POMO, EFormer, GASE, or L2D codebases. This design directly strengthens the neural comparison while keeping the claim limited to verified same-pipeline results.

TABLE II. NEURAL BASELINE SCOPE AND CONTROLLED SOLOMON RESULTS.

Method/scope	Run type	Feas.	D-gap	V-gap
Official AM	literature only	--	--	--
Official POMO	literature only	--	--	--
L2D style solvers	literature only	--	--	--
Official EFormer/GASE	literature only	--	--	--
AM style attention pointer	controlled	55/56	288.059	3.836
Edge aware Transformer	controlled	55/56	275.804	3.764
POMO style multistart	controlled	55/56	223.276	2.636
Relation-free Transformer	controlled	54/56	318.654	4.056
Proposed KG full	controlled	56/56	304.202	3.464

V. EXPERIMENTAL DESIGN

The experiments use the 56 original Solomon 100 customer instances and synthetic Solomon-style extensions containing clustered, random, and mixed customer distributions. Teacher routes are generated with OR-Tools, including repair runs for initially unresolved cases. Models are trained by imitation and evaluated by decoded feasibility, distance gap, and vehicle gap.

The synthetic stage provides a robustness view beyond the small original benchmark. Three independent 120 instance synthetic runs are summarized with mean values and 95 percent confidence intervals. Because OR-Tools is the teacher and reference, the reported gaps are conservative measures of how far the learned route construction remains from the solver output.

The evaluation uses four complementary indicators. Feasibility measures whether the decoded solution satisfies all route constraints. Vehicle gap measures how many additional vehicles the learned route uses relative to OR-Tools. Distance gap is the extra travel distance relative to OR-Tools. Maximum gap captures the worst-case deterioration. Together, these metrics distinguish constraint satisfaction, fleet size, average route quality, and tail risk.

The original Solomon instances are used to evaluate compatibility with a standard benchmark. The synthetic extensions are not intended to replace the benchmark; they provide repeated instance samples so that route selection and relation ablation can be tested across multiple seeds. This distinction is important because a framework claim should not rely only on a single small benchmark split.

VI. RESULTS AND ANALYSIS

Table II adds controlled neural baselines requested for the competitive context. The POMO style multistart baseline obtains the lowest Solomon mean distance gap among the local neural baselines (223.276) and a lower mean vehicle gap than the single policy models, while the AM style and edge aware baselines also improve route quality over the original relation-free Transformer. However, all three additional controlled baselines leave one Solomon instance infeasible. The proposed KG full model is therefore not claimed to dominate every quality metric; its main advantage in the original Solomon benchmark is complete decoded feasibility with explicit relation diagnostics.

TABLE III. ORIGINAL SOLOMON INSTANCE RESULTS

Method	Feas.	D-gap	Max gap	V-gap
Transformer baseline	54/56	318.654	621.406	4.056
KG-guided Transformer	56/56	304.202	683.267	3.464
KG-guided Transformer + repair	56/56	295.435	671.015	3.464

Table III reports the original Solomon results. KG guidance improves feasibility over the relation-free Transformer: the baseline decoded feasible routes for 54 of 56 instances, whereas the full KG-guided model decoded feasible routes for all 56 instances and reduced the mean vehicle gap. The repair variant further reduced the mean distance gap while retaining complete feasibility.

Table IV shows the Solomon relation ablation. Relation groups contribute differently: temporal and slack-oriented variants are competitive on the distance gap, while the full setting gives complete feasibility. The no slack setting has a low distance gap but loses one feasible instance, showing that relation selection changes the tradeoff between feasibility, distance, and vehicle count.

TABLE IV. SOLOMON RELATION ABLATION SUMMARY

Relation	Feas.	D-gap	V-gap	Relation specialization
spatial	55/56	295.447	3.455	spatial
temporal	56/56	272.667	4.268	temporal
capacity	55/56	276.861	3.964	capacity
feasible	55/56	295.612	3.909	feasible
slack	56/56	272.708	3.643	slack
no_slack	55/56	267.915	3.309	no_slack
full	56/56	304.202	3.464	full

Table V reports the synthetic medium route selection results, and Table VI summarizes the multi-seed synthetic robustness study. Objective-aware ensembling is stronger than any single learned model in terms of feasibility and distance gap. Across three 120 instance synthetic runs, the distance-oriented ensemble reaches a mean feasible rate of 0.997 and reduces the mean distance gap by 125.418 Solomon coordinate distance units relative to the relation-free Transformer. Table VII further shows that the hardest groups are R1 and RC1, where relation diversity and selection provide the largest feasibility improvements.

TABLE V. SYNTHETIC MEDIUM ROUTE SELECTION RESULTS

Method	Feas.	D-gap	V-gap
transformer_baseline	55/60	451.255	3.836
kg_ablation_full	55/60	404.900	3.218
kg_ablation_no_slack	59/60	445.001	3.153
adaptive_kg_transformer_l1	56/60	432.546	3.679
ensemble_all	60/60	344.374	1.867
ensemble_all_distance	60/60	311.390	2.633
router_ranker	57/60	437.149	3.263
router_nn	55/60	439.378	3.491

TABLE VI. MULTI-SEED SYNTHETIC ROBUSTNESS

Method	Feas. rate	D-gap	V-gap
transformer_baseline	0.858 +/- 0.062	427.827 +/- 22.013	4.557 +/- 0.960
kg_ablation_full	0.911 +/- 0.097	407.123 +/- 19.786	3.954 +/- 0.991
kg_ablation_no_slack	0.936 +/- 0.029	390.777 +/- 26.168	3.070 +/- 0.584
adaptive_kg_transformer_l1	0.914 +/- 0.082	393.310 +/- 13.080	3.769 +/- 1.181
router_nn	0.933 +/- 0.025	410.077 +/- 14.317	3.304 +/- 0.121
router_ranker	0.947 +/- 0.020	421.206 +/- 5.469	3.180 +/- 0.092
ensemble_all	0.997 +/- 0.005	336.183 +/- 9.882	2.031 +/- 0.099
ensemble_all_distanc	0.997 +/- 0.005	302.409 +/- 4.245	2.858 +/- 0.037

Table VIII reports paired t-tests across the three independent 120 instance synthetic runs. The distance-oriented ensemble significantly reduces the distance gap relative to the relation-free Transformer baseline (p=0.0059) and significantly increases the feasible rate (p=0.0472). The vehicle-oriented ensemble also significantly reduces the vehicle gap (p=0.0327). The full KG ablation improves the feasible rate, but the three-seed test does not show significance for that isolated comparison, which supports the interpretation that objective-aware route selection is the most robust component in the current experiments.

TABLE VII. HARD CLASS FEASIBILITY AND DISTANCE GAP

Group	Base feas.	Ens. feas.	Base gap	Ens. gap
C1	0.950	1.000	402.579	270.620
C2	1.000	1.000	343.069	211.234
R1	0.417	0.983	476.645	369.153
R2	1.000	1.000	464.693	323.813
RC1	0.783	1.000	485.504	334.981
RC2	1.000	1.000	430.212	305.588

TABLE VIII. PAIRED MULTI-SEED SIGNIFICANCE TESTS

Comparison	Metric	Delta	t	p
ensemble_all_distance vs baseline	distance gap reduction	125.418	12.979	0.0059
ensemble_all_distance vs baseline	feasible rate increase	0.139	4.437	0.0472
ensemble_all vs baseline	vehicle gap reduction	2.526	5.390	0.0327
kg_ablation_full vs baseline	feasible rate increase	0.053	1.142	0.3719

TABLE IX. INFERENCE TIME BENCHMARK ON SOLOMON INSTANCES

Model	Device	Repeats	Mean s/inst.	Min-Max s	Overhead
Relation-free Transformer	cuda	3	1.356	1.003-1.879	reference
Full KG-guided Transformer	cuda	3	1.668	1.259-2.152	+23.0%

The results also show that full relation information is not always the best distance-oriented configuration. On the original Solomon ablation, the no slack and slack variants produce lower mean distance gaps than the full relation model, while the full relation model gives complete feasibility. This suggests that relation richness can improve constraint awareness but may also change the model's route construction bias. Therefore, the ablation should be interpreted as evidence for controllability rather than as a simple ranking of relation sets.

The hard class analysis is especially informative. R1 and RC1 are more difficult because customer locations are less clustered and time windows are tighter. The relation free Transformer has a low feasible rate in R1, whereas the distance-oriented ensemble reaches near complete feasibility.

This indicates that route validation and candidate diversity are most valuable when a single learned policy is brittle.

VII. Discussion

The main finding is that explicit relation structure makes neural route construction more controllable. The KG is useful not only as an input feature source but also as an experimental mechanism for testing which routing knowledge supports feasibility and route quality. Objective-aware selection further shows that relation-specialized models can be combined without assuming that a single relation setting is best for every instance class.

The framework should be interpreted as a learned construction and diagnostic layer. It does not replace OR-Tools, which remains the reference solver and teacher. Instead, it provides a structured neural route generation pipeline that can be paired with local search, repair, or solver-based refinement in future work.

From a methodological perspective, the proposed KG has two roles. The first role is inductive bias: it exposes pairwise feasibility and routing structure to the Transformer before the model has to infer those relations from data. The second role is experimental transparency: relation groups can be added, removed, or selected, making it possible to explain performance differences in terms of routing knowledge rather than only model size or training loss.

From a practical perspective, the framework is most useful as a fast route construction component. A dispatcher or decision support system could use KG-guided decoding to generate feasible candidate routes quickly, then pass the best candidates to a local search, repair heuristic, or exact solver for improvement. In this setting, the neural model contributes speed and diversity, while the downstream optimizer contributes final route refinement decision support system could use KG-guided decoding to generate feasible candidate routes quickly, then pass the best candidates to a local search, repair heuristic, or exact solver for improvement. In this setting, the neural model contributes speed and diversity, while the downstream optimizer contributes final route refinement.

Table IX reports inference time on the RTX 3050 environment using the 56 Solomon instances and three repeats per instance. The relation-free Transformer averages 1.356 s per instance, while the full KG-guided Transformer averages 1.668 s per instance. Thus, explicit KG relation bias adds about 23.0 percent decoding overhead in this implementation, which is modest relative to the added feasibility and diagnostic benefits reported above.

VIII. Limitations and Future Work

The current model uses greedy decoding, and the learned routes remain worse than OR-Tools in terms of distance. The original Solomon set is also small, so synthetic extensions are useful for robustness but cannot replace additional external benchmarks. The added AM style, POMO style, and edge-aware Transformer baselines strengthen the same-pipeline comparison, but official reproduction of AM, POMO, L2D, EFormer, GASE, and other stronger neural or hybrid VRP solvers remains future work.

Another limitation is that the relation schema is manually designed from known VRPTW constraints. This is appropriate for a first framework study, but future research could learn relation weights or discover additional relation channels automatically. The router component also needs more data before it can consistently outperform simple objective aware ensembling. Finally, computational time should be reported more fully in future comparisons so that route quality can be interpreted together with inference cost.

IX. Conclusion

This study presented a multi-relation KG-guided Transformer framework for VRPTW. The KG is projected into Transformer attention as pairwise relation bias, and the framework is evaluated through ablation, adaptive variants, route selection, and multi-seed synthetic robustness. The results show improved feasibility over a relation-free Transformer and stronger robustness through objective-aware route selection. The contribution is a framework for operationalizing explicit routing knowledge inside neural attention and for analyzing how relation groups affect learned VRPTW construction.

References

[1] M. M. Solomon, "Algorithms for the vehicle routing and scheduling problems with time window constraints," Operations Research, vol. 35, no. 2, pp. 254-265, 1987, doi: https://doi.org/10.1287/opre.35.2.254.

[2] Google, "OR-Tools routing: vehicle routing," Google for Developers, Jan. 22, 2025. [Online]. Available: https://developers.google.com/optimization/routing. Accessed: May 17, 2026.

[3] Aigerim Bogyrbayeva; Meraryslan Meraliyev; Taukekhan Mustakhov; Bissenbay Dauletbayev, "Learning to Solve Vehicle Routing Problems: A Survey," arXiv (Cornell University), 2022. doi: https://doi.org/10.48550/arxiv.2205.02453.

[4] Xuan Wu; Di Wang; Lijie Wen; Yubin Xiao; Chunguo Wu; Yuesong Wu; Chaoyu Yu; Douglas L. Maskell; You Zhou, "Neural Combinatorial Optimization Algorithms for Solving Vehicle Routing Problems: A Comprehensive Survey with Perspectives," arXiv (Cornell University), 2024. doi: https://doi.org/10.48550/arxiv.2406.00415.

[5] K. T. Chung; C.K.M. Lee; Yung Po Tsang, "Neural combinatorial optimization with reinforcement learning in industrial engineering: a survey," Artificial Intelligence Review, 2025. doi: https://doi.org/10.1007/s10462-024-11045-1.

[6] Cheng Chi; Amine Mohamed Aboussalah; Elias B. Khalil; Juyoung Wang; Zoha Sherkat-Masoumi, "A Deep Reinforcement Learning Framework For Column Generation," arXiv (Cornell University), 2022. doi: https://doi.org/10.48550/arxiv.2206.02568.

[7] Jonas K. Falkner; Daniela Thyssens; Lars Schmidt-Thieme, "Large Neighborhood Search based on Neural Construction Heuristics," arXiv (Cornell University), 2022. doi: https://doi.org/10.48550/arxiv.2205.00772.

[8] Leo Baty; Kai Jungel; Patrick S. Klein; Axel Parmentier; Maximilian Schiffer, "Combinatorial Optimization enriched Machine Learning to solve the Dynamic Vehicle Routing Problem with Time Windows," arXiv (Cornell University), 2023. doi: https://doi.org/10.48550/arxiv.2304.00789.

[9] Jianjun Deng; Junjie Wang; Xiaojun Wang; Yiqiao Cai; Peizhong Liu, "Multi-Task Multi-Objective Evolutionary Search Based on Deep Reinforcement Learning for Multi-Objective Vehicle Routing Problems with Time Windows," Symmetry, 2024. doi: https://doi.org/10.3390/sym16081030.

[10] Rixin Wu; Ran Wang; Jie Hao; Qiang Wu; Ping Wang; Dusit Niyato, "Multiobjective Vehicle Routing Optimization With Time Windows: A Hybrid Approach Using Deep Reinforcement Learning and NSGA-II,"

IEEE Transactions on Intelligent Transportation Systems, 2024. doi: https://doi.org/10.1109/tits.2024.3515997.

[11] Fahong Yu; Meijia Chen; Xiaoyun Xia; Dongping Zhu; Qiang Peng; Kuibiao Deng, "Logistics Distribution Route Optimization With Time Windows Based on Multi-Agent Deep Reinforcement Learning," International Journal of Information Technologies and Systems Approach, 2024. doi: https://doi.org/10.4018/ijitsa.342084.

[12] Zixian Zhang; Geqi Qi; Wei Guan, "Coordinated multi-agent hierarchical deep reinforcement learning to solve multi-trip vehicle routing problems with soft time windows," IET Intelligent Transport Systems, 2023. doi: https://doi.org/10.1049/itr2.12394.

[13] Arash Mozhdehi; Mahdi Mohammadizadeh; Xin Wang, "Edge-DIRECT: A Deep Reinforcement Learning-based Method for Solving Heterogeneous Electric Vehicle Routing Problem with Time Window Constraints," arXiv (Cornell University), 2024. doi: https://doi.org/10.48550/arxiv.2407.01615.

[14] Dimeth Nouicer; Ikbal Chammakhi Msadaa; Khaled Grayaa, "A Novel Routing Solution for EV Fleets: A Real-World Case Study Leveraging Double DQNs and Graph-Structured Data to Solve the EVRPTW Problem," IEEE Access, 2023. doi: https://doi.org/10.1109/access.2023.3327324.

[15] Zefang Zong; Xia Tong; Meng Zheng; Yong Li, "Reinforcement Learning for Solving Multiple Vehicle Routing Problem with Time Window," ACM Transactions on Intelligent Systems and Technology, 2024. doi: https://doi.org/10.1145/3625232.

[16] Zengjian Yang, "Learning to Solve Vehicle Routing Problem with Soft Time Window via Collaborative Transformer," Online, 2024. doi: https://doi.org/10.65286/icic.v20i2.44243.

[17] Dian Meng; Zhiguang Cao; Yaoxin Wu; Yaqing Hou; Hongwei Ge; Qiang Zhang, "EFormer: An Effective Edge-based Transformer for Vehicle Routing Problems," Online, 2024. doi: https://doi.org/10.24963/ijcai.2024/954.

[18] Wei Li; Bing Tian Dai; Xueming Yan; Junying Zou; Zhijie Liang; Jingwen Li, "Learning to solving vehicle routing problems via local-global feature fusion transformer," Complex & Intelligent Systems, 2025. doi: https://doi.org/10.1007/s40747-025-02018-0.

[19] Zhenwei Wang; Ruibin Bai; Fazlullah Khan; Ender Ozcan; Tiehua Zhang, "Gase: graph attention sampling with edges fusion for solving vehicle routing problems," Memetic Computing, 2024. doi: https://doi.org/10.1007/s12293-024-00428-0.

[20] Cong Dao Tran; Thong Bach; Truong Son Hy, "Symmetry-preserving graph attention network to solve routing problems at multiple resolutions," arXiv (Cornell University), 2023. doi: https://doi.org/10.48550/arxiv.2310.15543.

[21] Jiwoo Son; Minsu Kim; Sanghyeok Choi; Hyeonah Kim; Jinkyoo Park, "Equity-Transformer: Solving NP-Hard Min-Max Routing Problems as Sequential Generation with Equity Context," Proceedings of the AAAI Conference on Artificial Intelligence, 2024. doi: https://doi.org/10.1609/aaai.v38i18.30007.

[22] Zhuoyi Lin; Yaoxin Wu; Bangjian Zhou; Zhiguang Cao; Wen Song; Yingqian Zhang; J. Senthilnath, "Cross-Problem Learning for Solving Vehicle Routing Problems," arXiv (Cornell University), 2024. doi: https://doi.org/10.48550/arxiv.2404.11677.

[23] Samuel J. K. Chin; Matthias Winkenbach; Akash Srivastava, "Learning to Deliver: a Foundation Model for the Montreal Capacitated Vehicle Routing Problem," arXiv (Cornell University), 2024. doi: https://doi.org/10.48550/arxiv.2403.00026.

[24] Jianan Zhou; Zhiguang Cao; Yaoxin Wu; Wen Song; Yining Ma; Jie Zhang; Chi Xu, "MVMoE: Multi-Task Vehicle Routing Solver with Mixture-of-Experts," arXiv (Cornell University), 2024. doi: https://doi.org/10.48550/arxiv.2405.01029.

[25] Federico Berto; Chuanbo Hua; Nayeli Gast Zepeda; Andre Hottung; Niels A. Wouda; Leon Lan; Park, Junyoung; Kevin Tierney; Jinkyoo Park, "RouteFinder: Towards Foundation Models for Vehicle Routing Problems," arXiv (Cornell University), 2024. doi: https://doi.org/10.48550/arxiv.2406.15007.

[26] Zefang Zong; Hansen Wang; Jingwei Wang; Meng Zheng; Yong Li, "RBG: Hierarchically Solving Large-Scale Routing Problems in Logistic Systems via Reinforcement Learning," Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. doi: https://doi.org/10.1145/3534678.3539037.

[27] Zhao Yun; Xiaoning Shen; Zhongpei Ge, "A Knowledge-Guided Multi-Objective Shuffled Frog Leaping Algorithm for Dynamic Multi-Depot Multi-Trip Vehicle Routing Problem," Symmetry, 2024. doi: https://doi.org/10.3390/sym16060697.

[28] Reza Shahbazian; Alessia Ciacco; Giusy Macrina; Francesca Guerriero, "Knowledge-guided hybrid deep reinforcement learning for the dynamic multi-depot electric vehicle routing problem," Computers & Operations Research, 2025. doi: https://doi.org/10.1016/j.cor.2025.107217.

[29] Jorin Dornemann, "Solving the capacitated vehicle routing problem with time windows via graph convolutional network assisted tree search and quantum-inspired computing," Frontiers in Applied Mathematics and Statistics, 2023. doi: https://doi.org/10.3389/fams.2023.1155356.

[30] Kassem Danach; Louai Saker; Hassan Harb, "Integrating Metaheuristics and Machine Learning for Enhanced Vehicle Routing: A Comparative Study of Hyperheuristic and VAE-Based Approaches," World Electric Vehicle Journal, 2025. doi: https://doi.org/10.3390/wevj16050258.

[31] Ying Zhou; Lingjing Kong; Hui Wang, "A multiobjective edge-based learning algorithm for the vehicle routing problem with time windows," Information Sciences, 2025. doi: https://doi.org/10.1016/j.ins.2025.122223.

[32] Conghui Wang; Zhiguang Cao; Yaoxin Wu; Long Teng; Guohua Wu, "Deep Reinforcement Learning for Solving Vehicle Routing Problems With Backhauls," IEEE Transactions on Neural Networks and Learning Systems, 2024. doi: https://doi.org/10.1109/tnnls.2024.3371781.

[33] S. Kosolsombat and C. Ratanavilisagul, "Modified ant colony optimization with selecting and elimination customer and re-initialization for VRPTW," Bulletin of Electrical Engineering and Informatics, vol. 11, no. 6, pp. 3471-3482, Dec. 2022, doi: https://doi.org/10.11591/eei.v11i6.3943.

[34] S. Kosolsombat and C. Ratanavilisagul, "Applied Deep Reinforcement Learning for Solving the Vehicle Routing Problem with Time Windows," in 2023 8th International Conference on Computational Intelligence and Applications (ICCIA), Haikou, China, Jun. 23-25, 2023, pp. 21-25, doi: https://doi.org/10.1109/ICCIA59741.2023.00012.

Deployment-Aware 30-Day Readmission Prediction in Resource-Limited Hospitals: Calibration, Threshold Policy, and Decision Utility

Samer Asad Malalha[1], Ma Burhanuddin[2*], Hatem T M Duhair[3], Jamil Abedalrahim Jamil Alsayaydeh[4], Mazen Farid[5]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[1, 2]
Department of Engineering Technology, Fakulti Teknologi Dan Kejuruteraan Elektronik Dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[3, 4]
Faculty of Information Science and Technology (FIST), Multimedia University, Melaka 75450, Malaysia[5]
Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia[5]

***Abstract*—Thirty-day hospital readmission is a well-established quality metric, and many clinical prediction models have been developed for this task; however, high discrimination does not by itself mean that a model is safe to use in discharge workflows. This study developed and applied an integrated deployment-oriented evaluation workflow in which calibration, threshold governance, and decision utility were treated as primary evaluation requirements rather than secondary diagnostics. Retrospective inpatient data collected between 2022 and 2024 from two resource-limited government hospitals were used (N = 30,000; readmission prevalence = 15.0%). The analysis was based on patient-level internal validation using non-overlapping training, validation, and held-out test partitions. A multilayer perceptron neural network and a random forest were evaluated using patient-level grouping (70% training, 15% validation, 15% test). Both models showed strong discrimination on the held-out test set (ROC-AUC = 0.868 for the neural network and 0.880 for the random forest), with nearly identical minority class detection (PR-AUC = 0.461 vs 0.460). However, calibration analyses separated the models despite identical Brier scores (0.095). The neural network showed lower expected calibration error (ECE = 0.013 vs 0.035) and near ideal probability scaling (calibration slope = 0.971, 95% CI: [0.94, 1.00]) compared with the random forest (slope = 1.202). Threshold analysis also showed that a default threshold could be unsafe, since recall was 0.244 at 0.50 but increased to 0.867 at 0.12, while false negatives dropped from 510 to 90. Decision Curve Analysis further supported the neural network, including a mean net benefit of 0.097 at a threshold of 0.12. Practically, threshold, model version, and monthly calibration summaries should be logged in an audit trail.**

Keywords*—Clinical AI deployment; hospital readmission prediction; expected calibration error; electronic health records; operating region; probability reliability; model governance

I. Introduction

The use of thirty-day hospital readmission as a measure of quality of care and as a target for predictive analytics in numerous health systems has been well-documented [1]-[4]. The increasing availability of electronic health records has allowed machine learning and deep learning models to process large amounts of high-dimensional clinical information [5], [6]. However, the field still faces a similar translation problem: models that perform well retrospectively do not always act safely or consistently when embedded in the workflow of discharge planning, care coordination, and follow-up allocation [7], [8].

Discrimination metrics, specifically the area under the receiver operating characteristic (ROC) curve, dominate model evaluation in the readmission prediction literature [9], [10]. While discrimination is valuable because it measures a model's ability to rank patients based on their risk, it does not necessarily imply that the model produces numerically valid predicted probabilities that can inform decisions regarding absolute risk [11], [12]. Therefore, methodological critiques have emphasized that clinical risk prediction needs to evaluate model performance using additional methodologies that assess the model's ability to produce valid predicted probabilities and that examine the decision-making consequences of model predictions in the context of realistic operating conditions [3], [12] – [15].

Calibration is central to this deployment question. Calibration occurs when the model predicts the actual rates of events; therefore, the predicted risk is expressed as a frequency statement in the population of interest [12], [13]. If calibration fails, then risk will be either systematically over- or under-predicted, and downstream decisions will potentially result in under-treatment, over-treatment, or misallocation of resources, regardless of whether the discrimination level of the model remains unchanged. These problems are not theoretical. Miscalibration can affect what patients receive: intensive discharge counseling, early outpatient follow-up, or post-discharge monitoring. As previously described, these interventions can directly affect patient safety and workload in environments where the margin is thin [8], [16]. Calibration monitoring has recently been added to the list of lifecycle management responsibilities through recent governance efforts, and is no longer viewed as a one-time diagnostic during model development [17], [18].

There exists a second type of deployment constraint related to converting predicted risk into actionable decisions. The current practice of setting a default threshold of .50 for binary classification in order to classify a patient as being at high risk of readmission is increasingly difficult to defend in clinical settings, since the choice of threshold rarely corresponds to clinical preference, the operational capacity of the institution, or

*Corresponding author.

the relative cost of different types of errors [7], [19]. Choice of threshold determines the balance between false negatives and false positives, and this balance has significant clinical implications for preventing readmissions. For example, missing a high-risk patient may prevent the triggering of a follow-up pathway that could have prevented the patient's deterioration, while flagging too many low-risk patients may lead to saturation of available staff capacity and reduce the focus on those most likely to benefit from follow-up. An operational way to address this issue is to document a threshold selection process based upon institutional capacity, e.g., select a threshold that would yield approximately 25 to 30 flagged discharges per week because that is the number of transitional care calls that a team can reliably make. During deployment, the selected threshold should be recorded along with the model version and the effective date so that changes in decision-making behavior can be tracked over time. This framing of threshold optimization as a clinical design decision instead of a statistical convention provides an opportunity to link clinical design to the technical capabilities of the model [20]-[22].

Decision Curve Analysis (DCA) provides a mechanism for evaluating whether action on a model improves outcome, across a range of clinically relevant thresholds, by explicitly incorporating the relative harms of false positives and false negatives into net benefit [23], [24]. Although DCA is increasingly being advocated for in guidelines for the evaluation and validation of prediction models [14], [25], it is relatively rare in the evaluation of hospital readmission prediction models, where evaluation typically focuses on discrimination and limited calibration metrics [3], [8]. This gap is important because a model can appear to be highly discriminative on ROC-AUC, yet fail to provide decision utility in the threshold ranges that are relevant to discharge workflows.

These limitations are accentuated in resource-constrained systems where staffing shortages, fragmented documentation, and limited follow-up capacity can amplify the effects of predictive error [2], [7]. In such settings, a deployment-ready readmission model should exhibit more than just discrimination. It should demonstrate stable calibration and predictable threshold behavior, so that the workflow consequences of acting on the model remain understandable and manageable.

Therefore, motivated by the need to evaluate readmission models in terms that are closer to clinical use, the current study develops and applies an integrated deployment-oriented evaluation workflow that combines discrimination, calibration reliability, threshold sensitivity analysis, and decision utility within a single operating-region assessment. The contribution of this work is not the introduction of a new classifier architecture. Rather, it is the operational integration of probability reliability, threshold behavior, and net clinical benefit into the evaluation of 30-day readmission prediction. Specifically, the study applies this workflow to retrospective multi-site inpatient data from resource-constrained hospitals and examines whether model outputs remain interpretable and actionable when translated into discharge-planning thresholds. By doing so, the manuscript positions readmission prediction as a decision-support problem, not only as a classification task evaluated by ROC-AUC [17], [18], [25].

The remainder of this study is organized as follows. Section II reviews related work on readmission prediction, calibration, threshold policy, and decision-curve-based evaluation in clinical AI. Section III describes the study design, data source, model development pipeline, and evaluation protocol, including calibration assessment, threshold sensitivity analysis, and decision curve analysis. Section IV presents the results across discrimination, calibration, threshold behavior, and decision utility. Section V interprets the findings in relation to deployment readiness, threshold governance, model monitoring, and study limitations. Section VI summarizes the main conclusions and implications for accountable readmission prediction in resource-limited hospital settings.

II. Related Works

A. Readmission Prediction and EHR-Based Machine Learning

Thirty-day readmission prediction has been studied extensively because readmission is both a quality indicator and an operational concern for hospitals. Earlier work has shown that electronic health records can support risk prediction by combining demographic information, laboratory values, vital signs, prior utilization, and admission-related variables [1], [5], [6]. Within this literature, machine learning and deep learning models have often been evaluated mainly by discrimination metrics, especially ROC-AUC, because these metrics summarize how well patients are ranked by predicted risk [6], [9]. This ranking perspective is useful, but it does not fully answer the deployment question faced by discharge teams: whether a predicted probability can be trusted when deciding which patient should receive follow-up, counselling, or post-discharge monitoring.

Systematic reviews of readmission prediction models have also shown that methodological quality and validation practices vary considerably across studies [2], [3]. In many cases, models report acceptable or high discrimination, but provide limited evidence about probability reliability, threshold behavior, and clinical usefulness under realistic operating constraints. This creates a gap between retrospective model performance and workflow-level applicability. The present study is positioned within this gap. Rather than treating readmission prediction as a purely ranking problem, it evaluates whether the model outputs remain usable when translated into probability-based discharge decisions.

B. Calibration and Probability Reliability in Clinical AI

Calibration has become increasingly important in clinical prediction because risk scores are often interpreted as probabilities during decision-making [12], [13]. A model with good discrimination may still overestimate or underestimate risk in clinically important strata. In readmission prediction, this distinction matters because a patient assigned a risk of 0.20 may be handled differently from a patient assigned a risk of 0.08, even if both patients are ranked correctly relative to others. When calibration is poor, the same threshold can route too many low-risk patients into follow-up or fail to identify patients whose risk is clinically meaningful.

Recent clinical AI guidance has therefore emphasized that deployment-oriented validation should not stop at ROC-AUC

[17], [18], [25], [26], [27]. Probability reliability, lifecycle monitoring, and recalibration planning are part of responsible model use, particularly when predictions are used to allocate finite clinical resources. In practical terms, this means that calibration should be monitored through measures such as reliability diagrams, Brier score, expected calibration error, and calibration slope. The current study follows this direction by comparing the neural network and random forest not only by discrimination, but also by the degree to which their predicted probabilities align with observed 30-day readmission rates.

C. Threshold Policy, Decision Utility, and Operating Regions

A second limitation in much of the readmission prediction literature is the treatment of classification thresholds as technical defaults rather than clinical policy choices. The conventional 0.50 threshold is rarely aligned with discharge-planning priorities, because hospitals may prefer to identify more high-risk patients even if this increases the number of false positives [7], [19], [32]. The appropriate threshold depends on the relative harm of missed readmissions, the available follow-up capacity, and the workflow used to act on model outputs. For example, if a transitional care team can make 30 follow-up calls per week, then the threshold should be chosen with that capacity in mind rather than inherited from a generic binary classifier.

Decision Curve Analysis provides a way to evaluate whether model-guided action produces clinical benefit across threshold values [23], [24]. This is especially relevant in resource-limited hospitals, where the cost of false positives is not only statistical but operational: each flagged discharge may require staff time, documentation, and follow-up coordination. By combining threshold sensitivity analysis with decision curve analysis, the present study moves beyond the search for a single "best" threshold and instead identifies a clinically acceptable operating region. This region-based approach links model evaluation with implementation planning because it specifies where recall, false negative reduction, probability reliability, and net benefit remain jointly defensible.

D. Positioning of the Present Study

The present study contributes to this literature by integrating four evaluation domains into a single deployment-oriented workflow: discrimination, calibration, threshold sensitivity, and decision utility. Its contribution is not the proposal of a new classifier architecture, nor does it claim that one model family is universally superior across settings. Instead, the study demonstrates how a readmission model can be examined before implementation by asking four linked questions: whether it ranks patients accurately, whether its probabilities are reliable, how its threshold choices affect missed readmissions and workload, and whether model-guided action provides positive net benefit. This positioning is especially relevant for resource-constrained hospitals, where a threshold decision has concrete operational consequences, for example, the number of follow-up calls that a discharge team can complete during a working week.

III. Methods

A. Study Design and Data Source

This study used a retrospective, multi-centre design based on real-world inpatient records collected between 2022 and 2024 from two government hospitals operating under resource-limited conditions. The dataset was drawn from Rafidia Hospital and Al-Watani Hospital, subject to institutional approval and de-identification procedures. These hospitals represent a clinically relevant resource-constrained setting because the records include heterogeneous documentation practices, non-trivial missingness patterns, and operational constraints that commonly affect clinical data quality in low-resource environments [2], [8]. Accordingly, the dataset was suitable for evaluating internal deployment readiness under non-ideal data conditions, but it was not used to establish external generalizability across all resource-limited hospitals. Fig. 1 illustrates the multidimensional structure to assess the clinical deployability and operating-region identification.

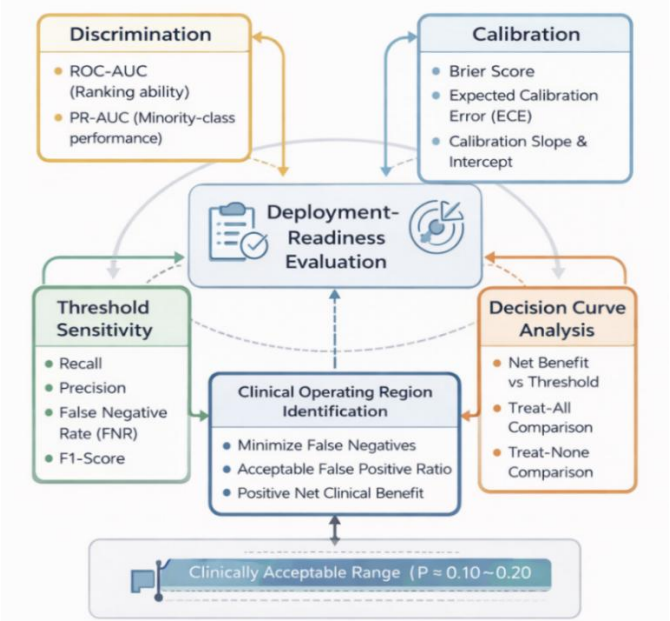


Fig. 1. Deployment-oriented clinical AI evaluation framework.

To minimise optimistic bias and prevent information leakage, we applied patient-level, group-aware data partitioning using unique patient identifiers (eid). A GroupShuffleSplit strategy was used to create non-overlapping training (70%), validation (15%), and held-out test (15%) subsets, with the constraint that no patient appeared in more than one subset. As a concrete operational safeguard, we programmatically verified that the intersection of patient identifiers across splits was empty and stored the split indices and random seed in an audit file so the exact partitions could be reproduced. All reported discrimination, calibration, Brier score, expected calibration error (ECE), and decision curve analysis (DCA) results were computed on the held-out test set only. Group-aware splitting is increasingly recommended in clinical machine learning to preserve validity and reduce inflated performance estimates that arise when repeated encounters from the same patient are inadvertently distributed across training and evaluation sets [8], [27]. This approach also aligns with contemporary guidance on validation practice for clinical prediction models [25].

No independent external validation cohort from a geographically or organizationally distinct institution was available for the present analysis. Therefore, all performance estimates should be interpreted as patient-level internal validation on a held-out test set rather than external validation. To reduce over-optimistic estimation within the available data, all preprocessing steps were fitted on the training set only, patient identifiers were kept non-overlapping across splits, and the final discrimination, calibration, threshold, and decision-curve results were calculated exclusively on the held-out test set.

The primary outcome was 30-day all-cause readmission following discharge. Analyses were conducted under institutional ethical approvals with de-identification procedures consistent with current reporting and governance expectations for clinical AI studies [17], [28]. See Table I.

TABLE I. COHORT CHARACTERISTICS AND DATA OVERVIEW

Category	Variable	Overall Cohort (N = 30,000)
Outcome distribution	30-day readmission, n (%)	4,500 (15.0%)
	No readmission, n (%)	25,500 (85.0%)
Demographics	Age, mean (SD), years	57.8 (18.4)
	Male sex, n (%)	15,960 (53.2%)
	Female sex, n (%)	14,040 (46.8%)
Admission characteristics	Length of stay, median (IQR), days	5 (3–8)
	Emergency admission, n (%)	18,750 (62.5%)
	Prior hospitalization (12 months), n (%)	9,840 (32.8%)
Feature composition	Demographic variables	6
	Laboratory variables	18
	Vital sign variables	6
	Admission-related variables	12
	Total features used in modeling	42
Missing data summary	Variables with <5% missingness	31 (73.8%)
	Variables with 5–20% missingness	9 (21.4%)
	Variables with >20% missingness	2 (4.8%)
	Imputation strategy	Median (continuous); Mode (categorical)
Data splitting strategy	Train set, n (%)	21,000 (70%)
	Validation set, n (%)	4,500 (15%)
	Test set, n (%)	4,500 (15%)
	Patient-level split	Applied
	Data leakage prevention	No patient overlap

B. Predictive Models

We evaluated two representative machine learning models: 1) a multilayer perceptron (MLP) neural network as the primary model, and 2) a random forest (RF) classifier as a baseline comparator. The purpose of this comparison was not to establish a full model leaderboard, but to examine how two commonly used nonlinear model families behave when evaluated through discrimination, calibration, threshold sensitivity, and decision utility under the same patient-level split and preprocessing protocol.

Neural network architecture and training. The MLP was implemented as a feed-forward network with three fully connected hidden layers (128, 64, and 32 units) using ReLU activations. Optimisation used Adam with an initial learning rate of 0.001 and L2 regularisation ($\alpha = 1 \times 10^{-4}$). Training proceeded for up to 300 iterations with early stopping based on validation performance, using a tolerance of 1×10^{-4} and patience of seven consecutive iterations without improvement. A fixed random seed (42) was used to support reproducibility.

A leakage-free, column-wise preprocessing pipeline was implemented for all data. Scaled, continuous variables were preprocessed with training set statistics only. Skewed continuous predictors (e.g., length of hospital stay and laboratory results) were standardized by median and inter-quartile ranges (IQR), whereas other continuous predictors were standardized by mean and standard deviation. Categorical predictors were treated equally during each split of the data, and missing data were replaced using medians for continuous variables and modes for categorical variables (see Table I). All preprocessing transformations were fit to the training set, and therefore, all subsequent transformations were performed without any additional modifications to the validation and testing sets, thereby eliminating potential feature-engineering-induced leakage from the training data. The Multilayer Perceptron (MLP) was chosen because it may be able to capture the non-linear relationships among the many, diverse demographic, laboratory, and admission-related variables common to EHR-based predictive tasks [1],[4],[6].

Random Forest Baseline Model. A Random Forest (RF) model was utilized as a baseline model because RF models are typically effective for structured tabular data, handle multiple types of predictor variables and have been extensively compared with other models within clinical predictive modeling applications [29] to allow comparison of the behavior of probabilities produced by both modelling families utilizing the same split of the data, same preprocessing pipeline, and same evaluation metrics. The objective of this research was not to achieve maximum discrimination via an exhaustive search of possible architectures and/or hyperparameters; rather, it was to evaluate how well the ML models were calibrated, sensitive to threshold decisions, and useful for making decisions within real-world operating conditions [18].

Benchmarking scope. Clinically established readmission scores such as LACE+ and HOSPITAL, as well as gradient-boosted tree models, are important comparators in broader readmission prediction research. However, the present study did not estimate these additional benchmarks. This decision reflects the study's primary aim, which was to demonstrate a deployment-oriented evaluation workflow rather than to optimize predictive performance across all possible model families. It also avoids introducing unreported model outputs or score reconstructions that were not part of the original analysis. Accordingly, the results should be interpreted as an internally validated comparison between NN and RF probability outputs, not as evidence that these two models exhaust the set of deployable alternatives for 30-day readmission prediction.

C. Multi-Dimensional Evaluation Framework

Consistent with current recommendations for evaluating prediction models intended for clinical use, assessment was structured across four complementary domains: discrimination, calibration, threshold sensitivity, and decision utility [3], [12], [17], [27]. This design reflects the position that discrimination alone is insufficient for clinical readiness because deployment decisions depend on probability validity and the consequences of acting on model outputs [25], [18].

Fig. 2 describes the model-training pipeline, probability generation, discrimination analysis, calibration assessment, threshold sensitivity evaluation, and decision curve analysis used to assess clinical reliability and operational suitability.

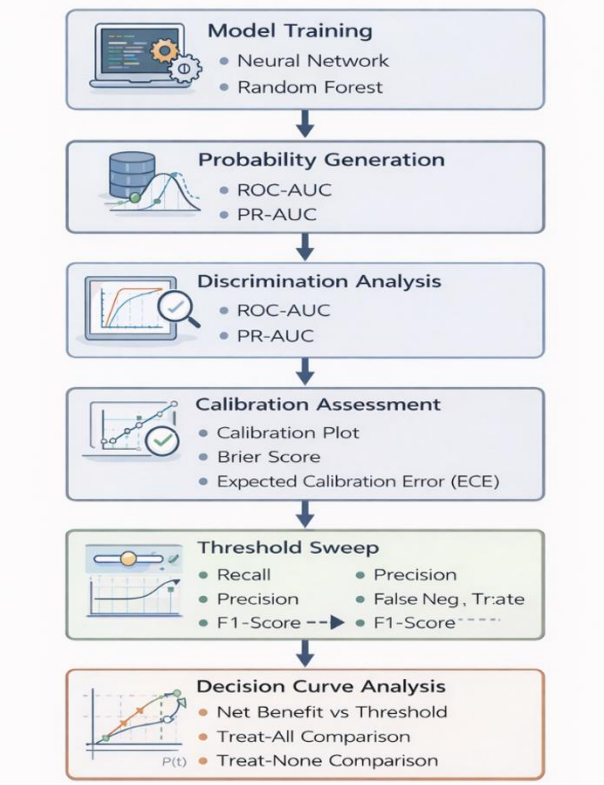


Fig. 2. Evaluation workflow for deployment-readiness assessment of clinical prediction models.

1) Discrimination assessment: Discrimination for both models was quantified using ROC-AUC, which summarises ranking performance across all possible thresholds. Because readmissions comprised approximately 15% of cases, we also evaluated precision–recall behaviour to characterise minority-class detection, and we report PR-AUC for the primary model as a complement to ROC-AUC under outcome imbalance [30], [31]. Discrimination was treated as necessary but not sufficient, and it was interpreted alongside calibration and decision-focused analyses.

2) Calibration assessment: Calibration analyses examined agreement between predicted probabilities and observed 30-day readmission rates. We computed reliability diagrams using observed versus predicted risk across deciles, the Brier score as an overall probabilistic accuracy measure, ECE as a bin-based summary of miscalibration, and calibration intercept and slope estimated via logistic recalibration [12], [27]. We also report the Hosmer–Lemeshow goodness-of-fit test as a conventional, though sample-size-sensitive, diagnostic. Calibration is repeatedly identified as a prerequisite for safe clinical deployment because it governs whether a predicted risk can be interpreted as an actionable probability rather than an uncalibrated score [12], [27]. Given outcome imbalance, we interpret the Brier score alongside complementary calibration measures, since prevalence can influence the absolute

magnitude of the Brier score [18]. ECE was computed using binned differences between predicted and observed event rates, consistent with common practice in probabilistic reliability evaluation [34]. For the recalibration sensitivity analysis, calibration metrics were reported separately for the original random forest, Platt-scaled random forest, and isotonic-regression-calibrated random forest, with all recalibration functions fitted on the validation set and evaluated only on the held-out test set. Reliability diagrams were generated by comparing observed and predicted 30-day readmission probabilities across risk deciles. These plots were used to visually assess departures from perfect calibration and are reported with the calibration results. Threshold Sensitivity Analysis.

Default probability thresholds, particularly 0.50, seldom reflect clinical risk tolerance or operational capacity in risk stratification tasks [7], [19], [32]. We therefore performed a systematic threshold sweep over probability cut-offs from 0.05 to 0.50 in increments of 0.01. At each threshold, we computed sensitivity (recall), precision, false negative rate (FNR), and F1-score (Table II). Because missed high-risk patients can undermine discharge safety pathways, the analysis emphasised FNR as a safety-relevant metric, while also documenting the corresponding workload implications captured by false positives [7], [34].

TABLE II. CONFUSION MATRIX PERFORMANCE METRICS AT DEFAULT (0.50) AND OPTIMIZED (0.12) PROBABILITY THRESHOLDS ON THE HELD-OUT TEST SET (N = 4,500)

Metric	Threshold = 0.50	Threshold = 0.12
True Positives (TP)	165	585
False Positives (FP)	151	1,018
True Negatives (TN)	3,674	2,807
False Negatives (FN)	510	90
Recall (Sensitivity)	0.244	0.867
Precision	0.522	0.365
F1-score	0.333	0.514
Overall Accuracy	0.853	0.754

3) Decision Curve Analysis (DCA): Clinical utility was evaluated using DCA across a range of probability thresholds [18], [19]. Net benefit was computed as:

$$\text{Net Benefit} = (TP / N) - (FP / N) \times (pt / (1 - pt)), \quad (1)$$

where, pt denotes the decision threshold. Model net benefit was compared against treat-all and treat-none strategies, and we also examined the default-threshold strategy to illustrate the impact of threshold choice on decision value. DCA provides an interpretable summary of whether decisions guided by model predictions improve outcomes under explicit trade-offs between false positives and false negatives, complementing discrimination and calibration analyses [24], [33]. Methodological guidance has increasingly positioned DCA as a core component of deployment-oriented validation, particularly when probability estimates are used to allocate finite interventions [18]. Fig. 3 illustrates net benefit across probability thresholds (0.05–0.50) compared with treat-all and treat-none strategies.

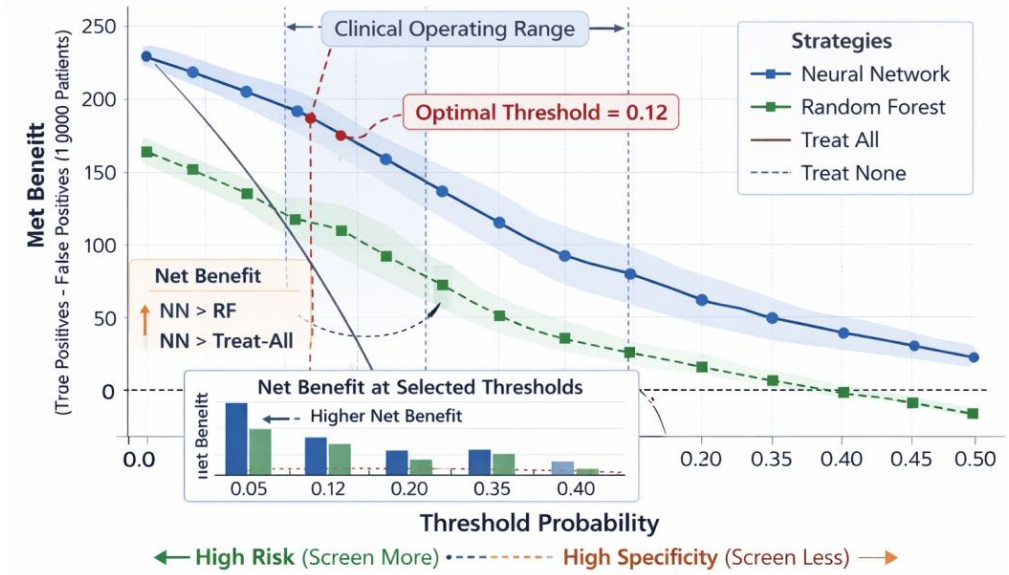


Fig. 3. Decision curve analysis of neural network and random forest models.

4) Clinical operating region identification: Rather than identify a "best" cut-off point for screening purposes based upon a single performance metric, we have identified a clinically acceptable operating region of thresholds which meet four criteria: improvement in the number of false negative results; an acceptable number of false positives to accommodate follow-up capacity, a positive net benefit in terms of DCA; and good calibration across all clinically relevant risk strata. The operating region approach is aligned with deployment-oriented expectations [16] [35] regarding consistency of decisions, safety, and fit into workflow under real-world constraints.

D. Statistical Analysis and Implementation

All analyses were implemented in Python using standard machine learning libraries within a reproducible computational environment. To quantify uncertainty, we used non-parametric bootstrap resampling on the held-out test set with 1,000 resamples drawn with replacement. For each resample, we recomputed ROC-AUC for both models and PR-AUC for the primary model, then recomputed calibration measures, including Brier score, ECE, calibration intercept, and calibration slope, and finally recomputed net benefit at selected thresholds for DCA. Confidence intervals were derived using the percentile method to obtain 95% intervals. This resampling-based approach provides a distributional view of metric variability, which is particularly relevant in imbalanced clinical prediction tasks and aligns with modern validation practice [23], [27]. Statistical significance was assessed at $\alpha = 0.05$ when hypothesis tests were reported, but interpretation prioritised effect magnitudes, interval estimates, probability reliability, and clinical interpretability over dichotomous p-value thresholds.

IV. Results

A. Discrimination Performance

Both models were highly discriminating at a global level on the test data for which they were partitioned using patient-level, group-aware partitioning. The Neural Network (NN) was able to achieve an ROC-AUC value of .868, showing that it ranked well in spite of high levels of imbalance. Using bootstrap resampling over 1000 iterations supports the stability of this estimate, and suggests that the ranking ability of the model was not due to a small number of samples in the test data.

A Random Forest (RF) Baseline performed better than the NN, with a higher ROC-AUC (.880), but also nearly identical precision – recall AUC (.460) for both models when looking at positive class detection, as readmission represented around 15% of the sample population. Overall, the discrimination results suggest that the models have similar risk rankings at a global level, although the RF has a slight advantage over the NN on ROC-AUC.

Discrimination, however, did not determine deployment readiness on its own. Because discharge workflows require binary decisions, and because these decisions depend on both threshold choice and probability reliability, subsequent analyses focused on threshold-dependent behaviour, calibration, and decision utility.

Fig. 4 shows the receiver operating characteristic curves used to summarize discrimination across thresholds. The precision–recall curve highlights positive-class detection under class imbalance, with readmissions accounting for approximately 15% of cases.

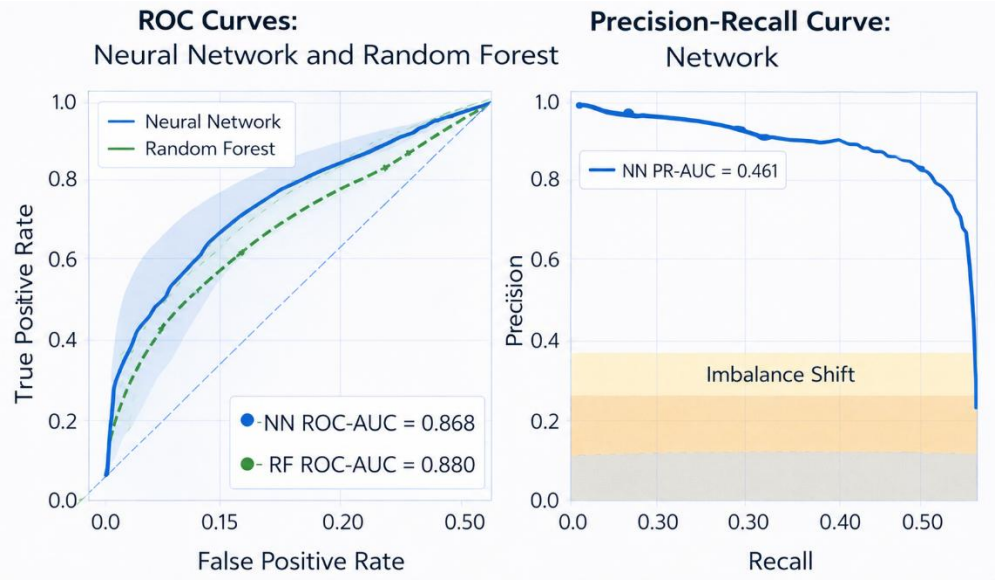


Fig. 4. ROC curves for the neural network and random forest models, and precision-recall curve for the neural network on the held-out test set.

Threshold-dependent behaviour showed why discrimination-only summaries were incomplete. At the default threshold of 0.50, recall for 30-day readmission was 0.244, meaning most readmissions were not flagged. When the decision threshold was set to 0.12 using a recall-oriented F-beta strategy ($\beta = 2$), recall increased to 0.867, with F1-score improving to 0.514 and overall accuracy equal to 0.754. This change reduced false negatives from 510 to 90 cases. It also changed operational workload: the number of flagged discharges increased from 316 at threshold 0.50 (TP + FP = 165

+ 151) to 1,603 at threshold 0.12 (TP + FP = 585 + 1,018), which corresponds to 7.0% versus 35.6% of the 4,500-patient test set (Table III). This illustrates that the same model can shift from under-detection to high-sensitivity screening depending on the operating threshold, with direct implications for follow-up capacity.

TABLE III. DISCRIMINATION AND CLASSIFICATION METRICS AT SELECTED THRESHOLDS (NEURAL NETWORK)

Metric	Threshold = 0.50 (Default)	Threshold = 0.12 (Optimized)
Recall (Sensitivity)	0.244	0.867
Precision	0.522	0.365
F1-score	0.333	0.514
Overall Accuracy	0.853	0.754
False Negatives (n)	510	90
False Negative Rate (FNR)	0.756	0.133

The comparison between thresholds 0.12 and 0.50, therefore, reflects a clinically meaningful trade-off. Lowering the threshold substantially reduced missed high-risk patients, while increasing false positives and the number of patients routed into a follow-up pathway. Although RF achieved marginally higher ROC-AUC, later sections show that probability reliability and decision-analytic performance differed between models, reinforcing that small differences in discrimination did not map cleanly onto deployment-relevant behaviour.

B. Calibration Assessment

Calibration analyses on the held-out test set examined whether predicted probabilities aligned with observed readmission frequencies. This matters because calibrated probabilities are interpretable as risk estimates, and interpretability is required for defensible threshold policies in discharge planning.

1) Brier score: A prevalence-only baseline that assigns every patient a 0.15 readmission risk achieved a Brier score of 0.1275. Both learned models improved substantially on this baseline. The NN produced a Brier score of 0.0950, and RF produced the same value, 0.0950, indicating comparable overall probabilistic accuracy when assessed by mean squared error of predicted probabilities. Bootstrap resampling supported the stability of these values. At the same time, the Brier score aggregates errors across the full probability range, so it can conceal differences in probability scaling and bin-level alignment, particularly in imbalanced datasets.

Probability density distributions of predicted risk scores for the neural network and random forest models evaluated exclusively on the patient-level held-out test set. Differences in spread and central tendency were examined further using calibration-specific metrics (see Fig. 5).

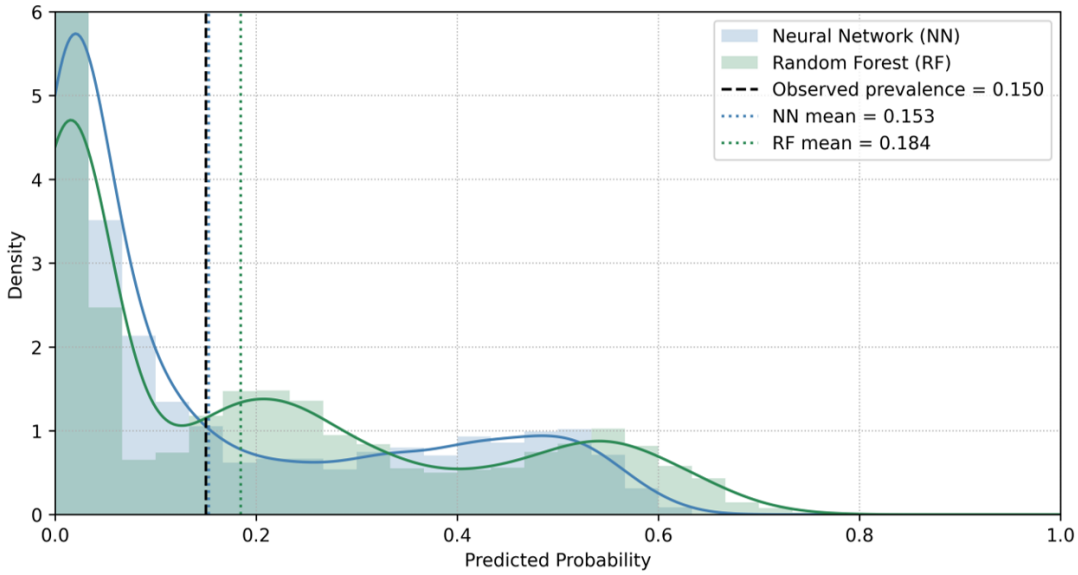


Fig. 5. Distribution of predicted 30-day readmission probabilities on the held-out test set.

2) Expected Calibration Error (ECE): Expected calibration error, computed on the held-out test set using equal-width bins, distinguished the models more clearly than the Brier score. The NN achieved ECE = 0.0128, whereas RF achieved ECE = 0.0348. The smaller value for NN indicates closer agreement between predicted probabilities and observed event rates across probability bins. In contrast, RF showed larger bin-level discrepancies, despite its marginal advantage in ROC-AUC. This contrast is important for clinical use because discharge decisions are often made in narrow probability bands, where bin-level miscalibration can shift which patients are flagged. Table IV shows the calibration analysis performed exclusively on the held-out test set. The results indicate that while both models have very similar total accuracy, the Neural Networks are capable of producing much more reliable probability estimates than Random Forests. Probability estimates are critical for clinical decision support, and therefore, this finding is highly relevant to our investigation.

TABLE IV. CALIBRATION PERFORMANCE ON THE HELD-OUT TEST SET

Metric	Neural Network (NN)	Random Forest (RF)	Ideal Value
Brier Score	0.095	0.095	0
Expected Calibration Error (ECE)	0.013	0.035	0
Calibration Intercept	−0.028	−0.216	0
Calibration Slope	0.971 (95% CI: 0.94–1.00)	1.202	1

3) Calibration intercept and slope: Recalibrating the logistic model to the held-out test data produced additional evidence as to how systematically biased and/or distorted the probability scales were for each model. The NN had an intercept of -0.028 and a slope of 0.971 (95% CI: 0.94-1.00). This demonstrates little systematic bias and appropriate probability scale distortion across risk levels. In contrast, the RF model's intercept was -0.216, and its slope was 1.202. This suggests that the RF model has larger deviations from ideal calibration and a larger probability scale distortion. Overall, these findings show that similar discrimination and Brier scores can conceal meaningful differences in uncalibrated probability reliability. The random forest showed larger deviation in ECE and calibration slope than the neural network, but this result should not be interpreted as evidence that the random forest model family is inherently less suitable for deployment. Rather, it indicates that the RF probabilities, as evaluated in their uncalibrated form, would require post-hoc recalibration assessment before any deployment-oriented comparison could be considered final.

This is evident in Fig. 6, where we show the observed versus the predicted 30-day readmission probability estimates for each risk level. The diagonal reference line represents perfect calibration, and the closer a model falls to this line, the more reliable its estimated probabilities are.

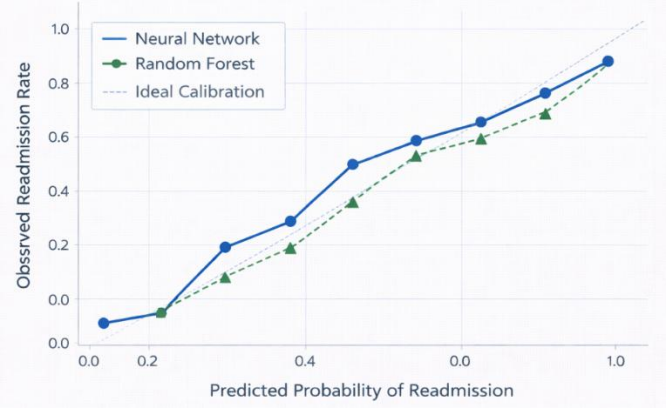


Fig. 6. Calibration curves (reliability diagrams) for neural network and random forest models.

C. Threshold Sensitivity Analysis

The results of the threshold sensitivity analyses are shown in Fig. 7. They indicate the size of the clinical performance variations associated with varying the decision threshold. For example, at the default threshold of 0.50, the model had a recall of 0.244. At the default threshold of 0.50, the model operated as a high-specificity but low-sensitivity rule. It identified only 24.4% of readmitted patients and therefore missed most patients who later experienced 30-day readmission. In contrast, when the threshold was set to 0.12, the recall increased to 0.867, the number of false negatives decreased from 510 to 90, which is roughly a sixfold decrease in missed readmissions. These data demonstrate that the conventional default thresholds used for decision-making in this context were inappropriate for achieving safe levels of care and further illustrate why the selection of the threshold should be viewed as a clinical decision based on available resources/workflow capacity rather than as an implicit statistical default.

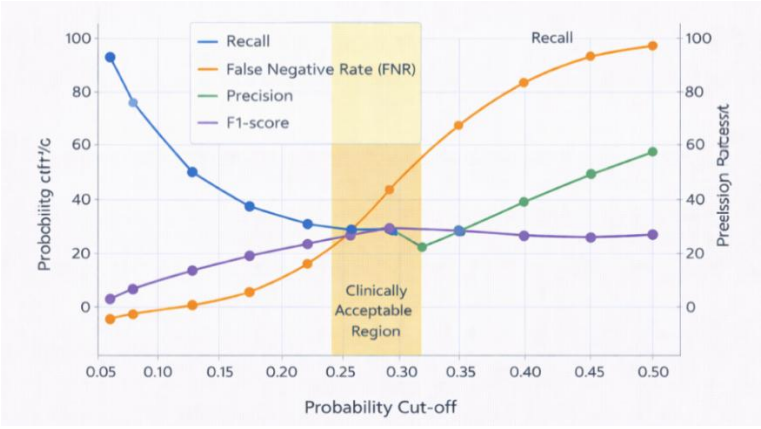


Fig. 7. Threshold sweep analysis for the neural network model.

Fig. 7 compares the recall, false negative rate (FNR), precision, and F1 score across different probability cut-off values ranging from 0.05 to 0.50. The shaded area (0.10–0.20) represents the clinically acceptable operating range, where there is high recall, a significant number of false negatives are eliminated, and sufficient decision utility is maintained.

D. Decision Curve Analysis

Decision Curve Analysis (DCA) examined whether taking action based on model predictions would result in net benefit over a wide range of thresholds. Across clinically relevant thresholds of 0.05 to 0.25, NN was found to have a consistent favorable net benefit when compared to treat-all and treat-none strategies. Using bootstrap resampling at an optimized threshold of 0.12, DCA produced a mean net benefit of 0.097 with a 95 percent confidence interval (CI): 0.093–0.101 that reflected relatively stable decision utility in the sensitivity-oriented operating range used for discharge screening.

Fig. 8 compares the net benefit across threshold probabilities for the NN strategy with the treat-all and treat-none strategies. NN has a favorable and relatively stable net benefit in the clinically relevant threshold ranges.

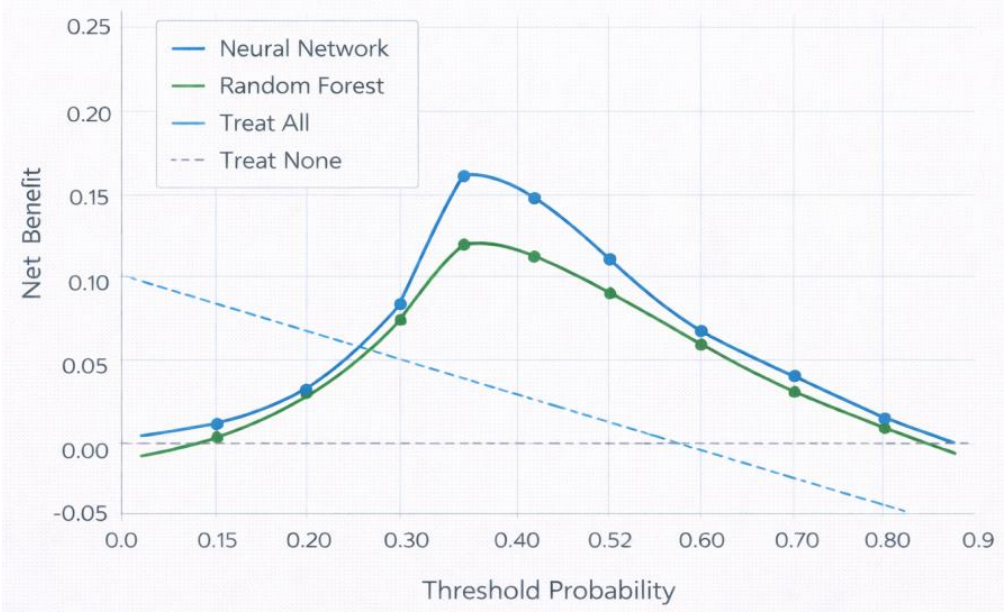


Fig. 8. Decision curve analysis comparing NN and RF strategies.

Although RF did show some positive net benefits at certain thresholds, it should be interpreted with caution, given the calibration distortions discussed previously. As discussed previously, these distortions may produce apparent shifts in gains across thresholds as a function of probability scaling distortion. Conversely, the combination of strong calibration and positive net benefits supports the suitability of NN for use in discharge planning workflows that rely on probability estimates to inform follow-up priority.

E. Clinical Operating Region Identification

The most appropriate deployment region, where the model would both correctly identify the individuals who needed treatment and provide enough information about how many people were treated, was determined by combining discrimination, calibration, threshold sensitivity, and decision utility instead of relying solely on one or more of those statistics to find the "best" threshold. The acceptable deployment region for clinical use was defined by four performance criteria: Recall should be at least 0.80; Calibration Error should be less than 0.02; Calibration Slope should be between 0.95 and 1.05 to allow the preservation of probability scaling; and there should be a positive Net Benefit relative to treating all and treating none. Using these criteria, thresholds ranging from 0.10 to 0.20 were found to be the best deployment region for NN. Within this range, NN provided good Recall (0.80 – 0.87), small Calibration Error (ECE = 0.013), and nearly perfect Probability Scaling (Slope = 0.971), with a positive Net Benefit. Therefore, the deployment region identified is a practical trade-off between significantly reducing False Negatives while maintaining Positive Decision Utility, see Table V.

TABLE V. INTEGRATED PERFORMANCE SUMMARY WITHIN THE CLINICALLY ACCEPTABLE OPERATING REGION (0.10–0.20)

Evaluation Dimension	Metric	Observed Performance (NN)	Interpretation Within Operating Region
Discrimination	ROC-AUC	0.868	Adequate global ranking performance
	PR-AUC	0.461	Stable minority-class detection under class imbalance
Calibration	Expected Calibration Error (ECE)	0.013	Minimal deviation between predicted and observed probabilities

	Calibration Slope	0.971 (95% CI: 0.94–1.00)	Probability scaling is closely aligned with the ideal (≈1)
Threshold Sensitivity	Recall	0.80–0.87	Substantial improvement in high-risk patient detection
	False Negatives	Reduced from 510 to 90 cases	Marked reduction in under-detection relative to default threshold (0.50)
Decision Utility	Net Benefit (DCA)	Positive vs. treat-all and treat-none strategies	Favorable clinical decision utility within operating range

V. Discussion

A. Beyond Discrimination: Why ROC-AUC Does Not Establish Deployment Readiness

This work supports the idea that it is possible for high discrimination and a number of potential deployment issues to be present together. The neural network produced a ROC-AUC of 0.868, while the random forest model generated a slightly better result at 0.880; however, neither of the global summary measures of rankings provided evidence that the probability predictions made by either model would be accurate enough to inform threshold-based discharge decisions. Discrimination, which is the focus of both of these models, addresses how well the model can rank patients in order of their predicted risk level, but does not provide information on the quantitative accuracy of the risk estimates made by the model to be used in actual decision-making processes (e.g., who should be discharged next) [12], [3], [27]. Therefore, there are numerous examples in the literature that have warned against using discrimination as the sole measure of clinical suitability for AI systems [12], [3], [27] and that other factors such as calibration stability, threshold alignment, and post-deployment monitoring and maintenance will need to be addressed to ensure trustworthy and lifecycle aware clinical AI systems [17], [16]. The findings reported here are consistent with this direction, suggesting that simply knowing the degree of discrimination of an AI system is not sufficient to determine if the system may be safely deployed and operated within real-world operational constraints.

B. Calibration Reliability Under Outcome Imbalance

The differences between the two predictive models were more apparent in calibration than in discrimination. Although both models produced the same Brier score (0.095) and similar discrimination, the calibration metrics indicated different levels of probability reliability under the present evaluation protocol. The uncalibrated random forest showed a higher expected calibration error (ECE = 0.035) and a calibration slope of 1.202, suggesting probability-scale distortion across risk strata. The neural network showed lower ECE (0.013) and a calibration slope closer to the ideal value (0.971; 95% CI: 0.94–1.00), indicating closer alignment between predicted and observed risk in this held-out test set. However, because post-hoc recalibration was not performed, these results should be interpreted as a comparison of the evaluated probability outputs, not as evidence that the random forest model family is inherently less suitable for deployment.

This distinction is particularly important when making clinical predictions using an imbalance of events. Global summaries such as the Brier Score pool squared error across all probability ranges, and since the event rate for this study was roughly 15%, the majority of errors in the decision-making space may have been diluted due to the large number of members in the non-event class. In contrast, ECE and the recalibration slope provide visibility to the reliability of the predictive probabilities in those bins/strata that are operationally relevant to the workflow, i.e., the decision thresholds [12], [18], [36].

Therefore, contemporary practice recommends assessing calibration as part of evaluating how well a model has been deployed, rather than treating it as simply another metric to report on discrimination performance [12], [3], [18]. The probability reliability of predictive modeling is not merely a cosmetic quality for hospital readmission workflows, but impacts who will be flagged, what clinicians interpret as the predictive risk means, and if the same threshold remains consistent in behavior over time.

C. Threshold Optimisation as a Clinical Design Decision

Beginning with Threshold Analysis, we see why threshold selection is considered part of both Clinical Design/Governance and not Conventional Practice. With a default threshold value of .50, we can see that Recall is .244, meaning there would have been a significant amount of readmission of patients that would have been missed using the Default Rule. Using an F-Beta Strategy (beta = 2) for a Recall-Oriented Approach, we can adjust the threshold to .12, and our Recall goes up to .867 while False Negatives drop to 90, down from 510.

The clinical implications are direct; a hospital discharge process that relies on identifying and addressing potential issues early in a patient's discharge will not perform as intended when the early identification tool misses three out of four readmissions.

The trade-off is also direct; to achieve a Recall of .867 using the Recall-Oriented Approach at the lower threshold (.12), we had to increase the number of discharges that were "flagged" to 1603 out of 4500 in the Test Set (True Positives + False Positives) compared to 316 at the higher threshold value of .50. This means that for every patient that has to undergo additional follow-up calls, and/or additional staffing capacity due to the need to schedule additional time for follow-up calls, the hospital will incur the associated cost. For this reason, recent literature recommends making the threshold selection based upon the specific trade-offs relevant to the deployment, documenting the reasoning behind the selection of the threshold, and matching the operating point of the system to the service capacity and level of risk that the organization can tolerate, rather than simply applying a generic default cut-off [25],[7],[37]. Our study provides empirical support for these recommendations: threshold selection affects the performance of systems that produce clinically important decisions; therefore, threshold

selection should be viewed as part of the design of a deployment, and thus should be subject to clearly defined escalation processes and audit trails for change control, rather than being treated as post-hoc tuning [38].

D. Decision Curve Analysis and Actionable Utility

Decision Curve Analysis (DCA) enhanced calibration and threshold sweep by asking a deployment-based question, i.e., does the model led to positive net benefits when using the model as a basis for action, across plausible thresholds [23] [24]. The Neural Network maintained favorable net benefits compared to the "treat all" and "treat none" strategies across the entire range of thresholds used for sensitivity-based screening of patients (i.e., approximately 0.05-0.25). Net Benefit was most stable at the optimal threshold of 0.12; Mean=0.097 (95% CI: 0.093-0.101) under bootstrap resampling. The stability of net benefits is important because it suggests that the observed decision advantage was not due to an isolated sample-specificity.

The importance of net benefit in this context stems from its ability to explicitly represent the trade-offs between false positives and false negatives through the threshold odds term pt/(1-pt) [24] [33]; and link those trade-offs to clinical actions [24] [33]. There is an increasing recognition of DCA as being complementary to calibration [25] [18] since a model can be well calibrated but still have limited utility in making decisions based on the model's output if the decision thresholds are not aligned with clinical practices; and similarly, a model may demonstrate high discrimination yet perform poorly in decision-making terms with respect to the ranges of thresholds that are clinically meaningful [23].

The inclusion of DCA with both calibration and threshold sensitivity allowed for a shift in how models were interpreted from global measures of performance to actionable outcomes, which are the appropriate standards of evaluation for settings where models will be deployed.

E. Implications for Resource-Constrained Healthcare Settings

The impact of prediction error on operations within resource-limited settings is amplified. Due to staffing limitations, variability in the practice of documenting, and limited follow-up options, false positives will continue to over saturate services while False Negatives will result in avoidable deterioration before receiving appropriate support [39]. As such, probability reliability becomes a necessity for workflows in resource-constrained settings. In addition, if predictions regarding patient risk are consistently skewed, clinicians may have difficulty developing consistent priorities based upon the recommendations of the clinical decision support system, and may lose confidence in the clinical decision support system's recommendations as they appear to contradict their own clinical impressions. Therefore, lifecycle-oriented clinical AI models address risk communication (calibrated), monitoring (ongoing), and governance processes that define drift and the need to recalibrate as normal occurrences, rather than as unusual events [17][16].

In addition, under the present uncalibrated evaluation protocol, the neural network showed favorable probability reliability and positive decision utility in the 0.10–0.20 operating region. This supports its use as the primary model for interpreting the current internal validation results. It should not, however, be read as a categorical deployment preference over the random forest, because RF recalibration using Platt scaling or isotonic regression was not tested. A practical implication of this approach is that threshold adjustments must be paired with lightweight monitoring. For example, a monthly report could document the current version of the model, the active threshold(s), the total number of discharges that were flagged by the model, the observed readmission rate among the flagged patients, and a decile-based calibration summary. This would allow for the tracking of threshold adjustments and performance drift through an audit trail. The incorporation of routine monitoring provides an ability to connect the behavior of the model to the operational decisions made as a result of those behaviors and allows for the safe iterative refinement of the model.

F. A Structured Framework for Deployment Readiness

Beyond evaluating models together, this study also provides a structure for evaluation to make deployability measurable. The central idea is to treat readiness as evidence across linked domains rather than as a single score. Discrimination produces a summary of ranking performance; calibration establishes whether the model's probability estimates are valid; threshold sensitivity describes how safety and workload will shift as policies change; and DCA quantifies whether an action improves outcomes in the range of thresholds that a workflow can sustain. The additional step of identifying a clinically acceptable operating region translates these metrics into a bounded set of thresholds where performance and utility are jointly defendable. Framed in this way, evaluation supports not only model selection but also planning for operation, including threshold governance, escalation rules, and monitoring triggers, which are required to deploy in safety-sensitive domains [40], [41]. While readmission prediction is focused upon here, the same logic applies to other imbalanced clinical tasks where small shifts in threshold policy can materially change patient routing.

G. Methodological Contribution

The research findings provide a deployment-focused interpretation of model performance, which may be summarized as three interrelated claims. Firstly, discrimination alone was not sufficient to establish deployability; the models were judged to have similar rankings for risk but different levels of probability reliability and decision utility. Secondly, having equal Brier scores did not mean the two models had the same level of calibration; both ECE and slope showed clinically important differences in how they scaled their probabilities. Finally, choosing the threshold to use changed the safety and workload implications of deploying the model, leading to threshold governance and decision-utility assessment being central to validating the model rather than optional additions to it. These findings collectively reframed model evaluation from "Is the model correct?" to "Do the probabilities generated by the model provide reliable information, does the decision policy used with the model behave consistently with expectations over the intended operating domain, and do decisions based upon the output of the model result in improved outcomes when the

model is deployed in a real-world environment subject to realistic constraints?".

H. Limitations and Future Directions

Several limitations should be considered when interpreting the findings of this study. First, although the dataset included 30,000 inpatient records from two resource-limited government hospitals, the evaluation remains retrospective and internally validated. No independent external cohort from a geographically or organizationally distinct hospital system was available. Therefore, the findings support the usefulness of the proposed evaluation workflow within the tested setting, but they should not be interpreted as evidence that the same model, threshold range, or calibration behavior will transfer unchanged to other resource-limited hospitals. External validation is required before broader implementation, ideally using a separate hospital network with different documentation practices, admission profiles, case mix, and follow-up capacity.

Second, the current evaluation assessed decision utility retrospectively rather than prospectively measuring the effect of model-guided intervention in routine discharge workflows. A prospective study would provide stronger evidence by examining operational outcomes after intervention allocation, such as follow-up completion rate, clinician adherence to risk alerts, readmission reduction, escalation decisions, and patient-level safety outcomes. For example, if a threshold of 0.12 routes approximately one third of discharged patients into a follow-up pathway, the clinical value of this threshold should be tested against the actual number of calls, appointments, or case-management reviews that staff can complete during a defined working week.

Third, the random forest model was evaluated without post-hoc recalibration using methods such as Platt scaling or isotonic regression. This limits the interpretation of the calibration comparison because part of the observed calibration gap may be correctable through recalibration. Future work should therefore evaluate whether recalibrated random forest probabilities change the model comparison in terms of ECE, calibration slope, threshold behavior, and net benefit.

Fourth, the comparative evaluation was limited to a neural network and a random forest. The study did not benchmark clinically established readmission scores such as LACE+ or HOSPITAL, nor did it include gradient-boosted tree models. This limits claims about comparative predictive superiority. Future studies should evaluate the same deployment-oriented workflow across a broader benchmark set, including clinical scores, random forest, neural networks, and gradient-boosted methods, using identical patient-level splits, preprocessing rules, calibration metrics, threshold sweeps, and decision curve analysis. This would allow the operating-region framework to be assessed as a general evaluation procedure rather than as a comparison restricted to two model families.

Finally, future studies should explicitly examine distribution shift and calibration drift after deployment. This requires continuous monitoring, predefined triggers for recalibration, and governance procedures for documenting changes in thresholds and model versions over time [16]. Such procedures would support the transition from one-time model validation to lifecycle management, which is necessary for responsible clinical AI implementation.

VI. Conclusions and Future Directions

This study shows that strong discrimination is necessary but not sufficient for evaluating clinical prediction models intended for discharge-planning decisions. ROC-AUC can indicate whether a model ranks patients correctly, but deployment also requires calibrated probabilities, transparent threshold policies, and evidence of decision utility within a clinically workable operating range.

The results illustrate this point clearly. Although the neural network and random forest showed broadly similar global performance, their calibration behavior differed. The random forest achieved slightly higher discrimination, but it also showed larger calibration error and greater probability-scale distortion. By contrast, the neural network produced more reliable probability estimates, behaved more predictably across threshold changes, and maintained positive net benefit within the clinically reasonable operating region. These findings indicate that practical model selection should consider probability reliability and threshold behavior, not only small gains in discrimination.

By combining discrimination, calibration assessment, threshold sensitivity analysis, and decision curve analysis, this study provides a deployment-oriented evaluation workflow for 30-day readmission prediction. In practical use, this means that the selected threshold should be linked to follow-up capacity, documented as a standing decision rule, and monitored through an audit trail that records the model version, active threshold, flagged-discharge volume, and monthly calibration summary.

These conclusions should be interpreted within the study design. The analysis used retrospective inpatient data from two resource-limited government hospitals and relied on patient-level internal validation with non-overlapping training, validation, and held-out test partitions. Therefore, the results support the value of the proposed workflow in the tested setting, but they do not establish that the same model, threshold range, calibration behavior, or model preference will transfer unchanged to other hospitals or to other benchmark models.

Future work should include independent external validation and prospective workflow evaluation. Important implementation outcomes include follow-up completion rate, clinician response to risk alerts, escalation decisions, readmission outcomes, staff workload, and calibration drift over time. Future studies should also evaluate recalibration strategies, monitor distribution shift, and define governance rules for updating thresholds and model versions during clinical use.

Authors' Contributions

The authors' contributions are as follows: "Conceptualization, Samer Asad Malalha; methodology, Samer Asad Malalha; software, Ma Burhanuddin; validation, Hatem T M Duhair; formal analysis, Jamil Abedalrahim Jamil Alsayaydeh; investigation, Samer Asad Malalha; resources, Mazen Farid; data curation, Hatem T M Duhair; writing—original draft preparation,, Jamil Abedalrahim Jamil Alsayaydeh, Hatem T M Duhair and Mazen Farid; writing—

review and editing, Ma Burhanuddin and Mazen Farid; visualization, Samer Asad Malalha; supervision, Ma Burhanuddin; project administration, Samer Asad Malalha; funding acquisition, Mazen Farid. All authors have read and agreed to the published version of the manuscript." The relevant terms are explained in the CRediT taxonomy.

ACKNOWLEDGMENT

The authors express their gratitude to the Centre for Research and Innovation Management (CRIM) at Universiti Teknikal Malaysia Melaka (UTeM) for their valuable support in this research.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was approved by the Clinical Research Center of An-Najah National University Hospital (Approval No. CRC_2025_0554; August 19, 2025). Approval to access anonymized patient records from government hospitals was also granted by the Palestinian Ministry of Health (Ref. No. 162/2161/2025; September 1, 2025), covering Rafidia Hospital and Al-Watani Hospital. All procedures were conducted in accordance with institutional and national ethical guidelines. Patient consent was waived due to the retrospective nature of the study and the use of fully anonymized electronic health records.

DATA AVAILABILITY

Data will be available on request.

REFERENCES

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[2] B. Van Grootven, P. Jepma, C. Rijpkema, L. Verweij, M. Leeflang, J. Daams, M. Deschodt, K. Milisen, J. Flamaing, and B. Buurman, "Prediction models for hospital readmissions in patients with heart disease: A systematic review and meta-analysis," BMJ Open, vol. 11, no. 8, e047576, Aug. 2021, doi: 10.1136/bmjopen-2020-047576.

[3] L. Wynants et al., "Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal," BMJ, vol. 369, m1328, Apr. 2020, doi: 10.1136/bmj.m1328.

[4] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, A. K. M. Z. Hossain, V. Shkarupylo, and J. Pusppanathan, "The experimental studies of the automatic control methods of magnetic separators performance by magnetic product," ARPN Journal of Engineering and Applied Sciences, vol. 15, no. 7, pp. 922-927, 2020.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[6] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063.

[7] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Medicine, vol. 17, no. 1, p. 195, Oct. 2019, doi: 10.1186/s12916-019-1426-2.

[8] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, no. 3, pp. 199–217, Mar. 2021, doi: 10.1038/s42256-021-00307-0.

[9] S.-M. Zhou, R. A. Lyons, M. A. Rahman, A. Holborow, and S. Brophy, "Predicting hospital readmission for campylobacteriosis from electronic health records: A machine learning and text mining perspective," Journal of Personalized Medicine, vol. 12, no. 1, p. 86, Jan. 2022, doi: 10.3390/jpm12010086.

[10] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, V. Shkarupylo, and D. A. P. P. Jkatisan, "Development of vehicle ignition using fingerprint," ARPN Journal of Engineering and Applied Sciences, vol. 14, no. 23, pp. 4045–4053, 2019. [Online]. Available: http://www.arpnjournals.org/jeas/research_papers/rp_2019/jeas_1219_8024.pdf.

[11] B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, and G. S. Collins, "Predictive analytics in health care: How can we know it works?" Journal of the American Medical Informatics Association, vol. 26, no. 12, pp. 1651–1654, Dec. 2019, doi: 10.1093/jamia/ocz130.

[12] M. N. Al-Andoli, Irianto, J. A. Alsayaydeh, I. M. Alwayle, C. K. N. Che Ku Mohd and F. Abuhoureyah, "Robust Overlapping Community Detection in Complex Networks With Graph Convolutional Networks and Fuzzy C-Means," in IEEE Access, vol. 12, pp. 70129-70145, 2024, doi: 10.1109/ACCESS.2024.3399883.

[13] R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell Jr., K. G. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: PART II, binary and time-to-event outcomes," Statistics in Medicine, vol. 38, no. 7, pp. 1276–1296, Mar. 2019, doi: 10.1002/sim.7992.

[14] C. G. Walsh, K. Sharman, and G. Hripcsak, "Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk," Journal of Biomedical Informatics, vol. 76, pp. 9–18, Dec. 2017, doi: 10.1016/j.jbi.2017.10.008.

[15] J. A. J. Alsayaydeh, W. A. Y. Khang, W. A. Indra, V. Shkarupylo and J. Jayasundar. 2019. Development of smart dustbin by using apps. ARPN Journal of Engineering and Applied Sciences. 14(21): 37033711.

[16] K. Lekadir et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, e081554, Feb. 2025, doi: 10.1136/bmj-2024-081554.

[17] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," BMJ, vol. 350, g7594, Jan. 2015, doi: 10.1136/bmj.g7594.

[18] B. Van Calster et al., "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance," The Lancet Digital Health, vol. 7, no. 12, p. 100916, Dec. 2025, doi: 10.1016/j.landig.2025.100916.

[19] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," Machine Learning, vol. 77, no. 1, pp. 103–123, Oct. 2009, doi: 10.1007/s10994-009-5119-5.

[20] K. B. Johnson et al., "Precision medicine, AI, and the future of personalized health care," Clinical and Translational Science, vol. 14, no. 1, pp. 86–93, Jan. 2021, doi: 10.1111/cts.12884.

[21] D. Piovani, R. Sokou, A. G. Tsantes, A. S. Vitello, and S. Bonovas, "Optimizing clinical decision making with decision curve analysis: Insights for clinical investigators," Healthcare, vol. 11, no. 16, p. 2244, Aug. 2023, doi: 10.3390/healthcare11162244.

[22] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo and C. Pellipus, "Homes appliances control using bluetooth," ARPN Journal of Engineering and Applied Sciences, vol. 14 (19), pp. 3344-3357, 2019.

[23] J. Benitez-Aurioles, L. Wynants, N. Peek, P. Goodley, P. Crosbie, and M. Sperrin, "The continuous net benefit: Assessing the clinical utility of prediction models when informing a continuum of decisions," Diagnostic and Prognostic Research, vol. 10, no. 1, p. 8, 2026, doi: 10.1186/s41512-026-00224-z.

[24] A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests," BMJ, vol. 352, i6, Jan. 2016, doi: 10.1136/bmj.i6.

[25] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, "Developing clinical prediction models: A step-by-step guide," BMJ, vol. 386, e078276, 2024, doi: 10.1136/bmj-2023-078276.

[26] A. Subbaswamy and S. Saria, "From development to deployment: Dataset shift, causality, and shift-stable models in health AI," Biostatistics, p. kxz041, 2019, doi: 10.1093/biostatistics/kxz041.

[27] R. D. Riley et al., “Calculating the sample size required for developing a clinical prediction model,” BMJ, p. m441, 2020, doi: 10.1136/bmj.m441.

[28] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, 1st ed. Geneva, Switzerland: World Health Organization, 2021.

[29] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[30] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, “The receiver operating characteristic curve accurately assesses imbalanced datasets,” Patterns, vol. 5, no. 6, p. 100994, Jun. 2024, doi: 10.1016/j.patter.2024.100994.

[31] J. A. J. Alsayaydeh, Irianto, A. Aziz, C. K. Xin, A. K. M. Z. Hossain and S. G. Herawan, “Face Recognition System Design and Implementation using Neural Networks” International Journal of Advanced Computer Science and Applications(IJACSA), 13(6), June 2022, pp. 519-526. http://dx.doi.org/10.14569/IJACSA.2022.0130663.

[32] M. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, United Kingdom: Oxford University Press, 2003.

[33] M. J. Pencina, R. B. D’Agostino, and E. W. Steyerberg, “Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers,” Statistics in Medicine, vol. 30, no. 1, pp. 11–21, Jan. 2011, doi: 10.1002/sim.4085.

[34] F. Futami and M. Fujisawa, “Information-theoretic generalization analysis for expected calibration error,” arXiv:2405.15709, 2025, doi: 10.48550/arXiv.2405.15709.

[35] M. Imani, M. Joudaki, A. Bagheri, and H. R. Arabnia, “Why ROC-AUC is misleading for highly imbalanced data: In-depth evaluation of MCC, F2-score, H-measure, and AUC-based metrics across diverse classifiers,” Technologies, vol. 14, no. 1, p. 54, 2026, doi: 10.3390/technologies14010054.

[36] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.

[37] M. P. Sendak et al., “Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study,” JMIR Medical Informatics, vol. 8, no. 7, e15182, Jul. 2020, doi: 10.2196/15182.

[38] J. Feng et al., “Clinical artificial intelligence quality improvement: Toward continual monitoring and updating of AI algorithms in healthcare,” npj Digital Medicine, vol. 5, no. 1, p. 66, May 2022, doi: 10.1038/s41746-022-00611-y.

[39] S. E. Davis, R. A. Greevy, T. A. Lasko, C. G. Walsh, and M. E. Matheny, “Detection of calibration drift in clinical prediction models to inform model updating,” Journal of Biomedical Informatics, vol. 112, p. 103611, Dec. 2020, doi: 10.1016/j.jbi.2020.103611.

[40] J. A. J. Alsayaydeh, Irianto, M. F. Ali, M. N. M. Al-Andoli and S. G. Herawan, "Improving the Robustness of IoT-Powered Smart City Applications Through Service-Reliant Application Authentication Technique," in IEEE Access, vol. 12, pp. 19405-19417, 2024, doi: 10.1109/ACCESS.2024.3361407.

[41] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” Science, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: 10.1126/science.aaw4399.

Expert Evaluation of the Proposed ERP Model for Legal Compliance and Adaptation

Shkëlqim Miftari, Azir Aliu, Artan Luma
South East European University, Tetovo, North Macedonia

***Abstract*—This study presents an expert-based evaluation of a proposed context-aware ERP model designed to support adaptive legal compliance in multilingual regulatory environments, with a focus on North Macedonia. The model integrates modular ERP architecture, AI-driven legal reasoning, multilingual natural language processing, adaptive learning, and human oversight and is intended to support continuous alignment with changing legislation. A structured questionnaire was administered to 51 experts from fields, including ERP systems, software engineering, artificial intelligence, legal compliance, and public administration, to assess the model's conceptual clarity, technical feasibility, multilingual support, scalability, and compliance functionality. The findings provide expert-based conceptual support for the proposed framework, particularly regarding its relevance, architectural coherence, and potential applicability, but they should not be interpreted as evidence of operational effectiveness in a deployed ERP environment. Participants particularly emphasized the importance of multilingual support and human validation mechanisms for maintaining accountability and trust in AI-assisted compliance processes. The evaluation also identified implementation challenges related to integration complexity, AI interpretability, and organizational readiness, highlighting areas for future prototype development and empirical validation.**

Keywords*—ERP systems; legal compliance; artificial intelligence; multilingual NLP; adaptive learning; expert evaluation; compliance automation

I. Introduction

Enterprise Resource Planning (ERP) systems face increasing pressure to adapt to rapidly evolving legal environments, particularly in jurisdictions characterized by complex and multilingual regulatory frameworks. In North Macedonia, where legislative updates are frequent and multilingualism is institutionally embedded, traditional ERP systems struggle to maintain legal compliance efficiently. Static compliance mechanisms configured during system design are often insufficient for responding to dynamic legal conditions, resulting in operational inefficiencies and increased compliance risks.

Although recent ERP research has explored artificial intelligence (AI) and process automation, limited attention has been given to integrated models that combine AI-driven legal reasoning, multilingual natural language processing (NLP), and adaptive learning within a unified compliance framework. This limitation is especially evident in regions such as the Western Balkans, where linguistic diversity and fragmented digital infrastructure further complicate interoperability and legal adaptability.

To address these challenges, this study proposes a modular and context-aware ERP model that integrates AI-based legal compliance mechanisms, bilingual support in Macedonian and Albanian, and adaptive learning for continuous regulatory alignment. The model is structured around a four-layer compliance framework consisting of a Legal Data Interface, an AI Reasoning and Adaptation Engine, a Human Oversight Layer, and an ERP Integration Module. Together, these components are intended to support explainable and traceable compliance adaptation while minimizing disruption to business operations.

The contribution of this study lies in the conceptual integration of ERP compliance adaptation, AI-assisted legal reasoning, multilingual Macedonian-Albanian NLP support, adaptive learning, and human oversight within a unified four-layer framework. Unlike traditional ERP systems that depend on static compliance rules and manual updates, the proposed model treats legal compliance as a continuous and reviewable adaptation process. Unlike general AI-assisted ERP systems, the focus is specifically on legally accountable compliance adaptation rather than broad process automation. Unlike generic legal NLP systems, the proposed framework connects legal interpretation to ERP configuration and workflow adaptation. Unlike conventional rule-based compliance modules, it emphasizes traceability, expert validation, and adaptability in legally dynamic and multilingual environments such as North Macedonia.

The primary objective of this study is to evaluate the model's conceptual soundness, technical feasibility, scalability, and contextual relevance through expert-based theoretical validation. A structured questionnaire was administered to professionals in ERP systems, artificial intelligence, legal compliance, public administration, and software engineering to assess the proposed framework's applicability in legally dynamic environments.

The remainder of the study is organized as follows: Section II reviews related research on ERP adaptation, AI-driven compliance, and multilingual systems. Section III presents the proposed ERP architecture. Section IV describes the research methodology, while Section V presents the evaluation results. Section VI discusses the findings, followed by limitations in Section VII and conclusions with future work directions in Section VIII.

II. Background and Related Work

Enterprise Resource Planning (ERP) systems are increasingly required to operate in dynamic regulatory and

multilingual environments, where traditional static architectures are insufficient for maintaining long-term legal compliance and operational flexibility. Contemporary research has therefore shifted toward adaptive ERP models capable of integrating artificial intelligence (AI), natural language processing (NLP), and context-aware automation to support evolving organizational and regulatory demands.

A. ERP Adoption and Legal Compliance

ERP systems are increasingly shifting from rigid customization models toward adaptive architectures capable of supporting legal compliance in dynamic regulatory environments. Traditional ERP customization often relies on hardcoded extensions that reduce maintainability and scalability when regulations change frequently. In contrast, adaptive frameworks promote modular and constraint-aware flexibility, enabling business processes to evolve while preserving structural integrity and execution safety.

The ADEPT2 framework exemplifies this transition by supporting runtime process adaptation through controlled operations such as fragment insertion and deletion while validating modifications against compliance constraints [8]. Complementary multi-time binding mechanisms further enable process variants and data structures to be configured at either design time or runtime, reducing the limitations associated with static one-size-fits-all ERP customization approaches [8].

Adaptation is particularly important in multi-jurisdictional contexts where ERP systems must accommodate varying tax regulations, data protection requirements, and compliance mandates. AI-based approaches have also been examined in tax planning and regulatory compliance contexts, indicating their potential relevance for structuring compliance-related information and supporting automated verification and governance processes [4]. Similarly, large language model approaches have demonstrated potential for converting legal documents into machine-readable governance policies applicable to regulated sectors affected by overlapping legal frameworks such as GDPR and Basel III [14].

Static compliance mechanisms configured during system design remain limited in their ability to respond to evolving legal interpretations and runtime operational changes. Research on adaptive compliance, therefore, increasingly emphasizes predictive and continuous monitoring approaches capable of identifying potential violations before they occur [1]. These approaches also highlight the importance of explainability, periodic model retraining, and human oversight to ensure trustworthy and legally accountable AI-assisted compliance processes [4].

As ERP systems continue to evolve, adaptive compliance capabilities such as automated regulatory tracking, explainable AI reasoning, and context-aware rule management are becoming essential for maintaining operational flexibility and legal alignment across diverse regulatory environments.

B. AI in ERP Systems

Artificial intelligence is increasingly integrated into ERP systems to enhance decision support, adaptability, and compliance management in complex organizational environments. Rather than replacing core ERP functions, AI augments operational processes through predictive analytics, anomaly detection, contextual automation, and intelligent resource optimization. AI-supported ERP systems can improve decision-making across finance, logistics, human resources, and customer management while operating within predefined organizational and regulatory constraints [6].

AI also reduces cognitive overload by filtering large ERP datasets to identify patterns, risks, and anomalies that may not be easily detectable through manual analysis. Applications include fraud detection, predictive maintenance, demand forecasting, and workforce analytics, all of which contribute to improved responsiveness and operational efficiency [5].

However, the integration of AI into ERP environments introduces important governance and compliance challenges. Since ERP operations frequently intersect with regulated domains such as taxation, labor management, and financial reporting, AI-generated outputs must remain explainable, traceable, and legally accountable [16]. To address these concerns, human-in-the-loop mechanisms are increasingly incorporated into ERP-AI architectures, allowing ambiguous or high-impact decisions to be reviewed and validated by human operators [6]. Explanation interfaces, audit trails, and override capabilities, therefore, become essential components for maintaining institutional trust and regulatory compliance.

Contemporary ERP research also highlights a transition from traditional automation toward bounded process autonomization, where AI systems assume situational control over workflows while remaining constrained by explicit operational, ethical, and governance rules [15]. Within this context, responsible AI integration enables ERP systems to evolve from static transactional platforms into adaptive decision-support ecosystems capable of balancing operational efficiency with regulatory accountability.

These studies show that AI can extend ERP systems beyond static transactional processing toward intelligent decision support. However, much of the existing AI-in-ERP literature focuses on operational optimization, forecasting, anomaly detection, and process automation rather than legal compliance adaptation. The proposed model builds on this direction by focusing specifically on AI-assisted interpretation of legal changes, compliance recommendation generation, and human validation within ERP workflows.

C. Multilingual ERP Systems

The deployment of ERP systems in multilingual environments introduces significant challenges related to linguistic diversity, semantic ambiguity, and interoperability. Within such settings, NLP functions as a critical bridge between structured enterprise data and human communication, enabling applications such as document processing, compliance monitoring, automated translation, and conversational interfaces. However, the effectiveness of multilingual ERP systems depends heavily on the availability of reliable linguistic resources and domain-specific language models.

These challenges are particularly evident in the Western Balkans, where underdeveloped NLP infrastructures and limited access to AI technologies hinder enterprise system adaptability and interoperability [7]. Research has shown that NLP tools for

Macedonian and Albanian remain comparatively under-resourced, limiting accurate processing of legal and organizational texts in multilingual ERP environments [9], [13]. Similar limitations have also been identified across South Slavic languages due to insufficient corpora and weak language-processing ecosystems [10].

Within ERP contexts, NLP technologies have demonstrated strong potential for improving automation, user interaction, and knowledge extraction through functions such as invoice parsing, email classification, and semantic process interpretation [17]. AI-augmented business process management research further emphasizes that multilingual natural language understanding is essential for adaptive and explainable ERP environments operating across diverse user groups and regulatory contexts [3], [6].

In the context of North Macedonia, multilingual support represents both a technical and institutional requirement. The integration of Macedonian and Albanian language processing is essential for ensuring inclusive system interaction, accurate compliance interpretation, and operational consistency across public and private sector environments. Consequently, regional digital transformation efforts increasingly emphasize investment in multilingual AI infrastructures, localized NLP services, and language-aware enterprise technologies capable of supporting legally and linguistically diverse organizational ecosystems [2].

Overall, the reviewed literature shows increasing interest in adaptive ERP architectures, AI-supported process automation, and multilingual NLP. However, these research streams are often treated separately. Existing ERP adaptation studies emphasize process flexibility, AI-in-ERP studies often focus on operational optimization, and multilingual NLP studies address language-processing limitations without necessarily embedding them into ERP compliance workflows. This study responds to that gap by proposing a conceptual ERP compliance framework that combines legal-data interpretation, AI-assisted reasoning, multilingual Macedonian-Albanian support, adaptive learning, and human validation within a unified four-layer architecture.

III. Overview of the Proposed ERP Model

The proposed model builds on a modular client-server ERP system developed for the North Macedonian enterprise context using PHP, Laravel, and the Model-View-Controller (MVC) architectural pattern. The core ERP system separates presentation, business logic, and data management layers and supports functional domains such as Finance, Human Resources, and Inventory Management. In this study, the ERP system is extended conceptually through an external AI-driven compliance framework designed to support adaptive legal compliance in multilingual regulatory environments.

A central feature of the proposed model is a dedicated four-layer compliance framework consisting of a Legal Data Interface, an external AI Reasoning and Adaptation Engine, a Human Oversight Layer, and an ERP Integration Module. This framework is designed to operate independently from transactional ERP services, enabling legal adaptation recommendations to be generated, reviewed, and transferred to the ERP system without disrupting core business operations.

The architecture integrates AI-based legal reasoning, multilingual natural language processing, adaptive learning mechanisms, and human validation procedures to support explainable and traceable compliance management. Through this combination of modular ERP functionality and intelligent compliance services, the model aims to provide a scalable and context-aware solution for enterprise environments characterized by evolving legal requirements and multilingual interaction.

A. Architectural Overview

The ERP system is designed as a modular and context-aware architecture that supports both enterprise operations and adaptive legal compliance. The core ERP system was developed using PHP, Laravel, and the Model-View-Controller (MVC) pattern, separating data access, business logic, and presentation layers to improve maintainability, scalability, and flexibility [11]. Functional domains such as Human Resources, Finance, and Inventory Management operate as independent ERP modules that can be deployed and customized according to organizational requirements.

The proposed compliance architecture is positioned outside the transactional ERP core and communicates with it through the ERP Integration Module. This separation allows the AI-based compliance components to operate as external decision-support and recommendation services rather than as embedded transactional ERP functions [12].

The architecture is intended to use a relational MySQL database with normalized schemas and indexed queries to ensure data consistency, efficient access, and scalable transaction management [11]. A defining characteristic of the system is the separation between operational ERP services and compliance management through a four-layer adaptive framework consisting of the Legal Data Interface, AI Reasoning and Adaptation Engine, Human Oversight Layer, and ERP Integration Module [12].

The Legal Data Interface is intended to collect and structure legislative information from official sources using NLP techniques that transform unstructured legal texts into machine-interpretable representations. The AI Reasoning and Adaptation Engine is designed to apply ontology-based reasoning, rule inference, and reinforcement learning to determine how legislative modifications could affect ERP workflows. The proposed system would identify potentially impacted business domains and generate recommendations for process adaptation, such as revised payroll calculations or compliance-reporting procedures. Reviewer feedback and organizational data could be used in future implementation to refine interpretive accuracy over time [12].

To ensure accountability and institutional control, AI-generated recommendations would be reviewed within the Human Oversight Layer, where authorized experts could validate, reject, or modify proposed changes. Each action would be documented through metadata containing timestamps, legal references, and reviewer information to support transparency and traceability. Approved adaptations would then be transferred to the ERP Integration Module, which is intended to

apply configuration updates through secure APIs without requiring full system redeployment [12].

This separation between compliance services and transactional business logic enables continuous regulatory adaptation while maintaining operational stability. By integrating semantic reasoning, adaptive learning, multilingual processing, and human validation, the architecture supports legally accountable and scalable ERP environments suited to dynamically evolving regulatory contexts.

B. AI-Driven Legal Compliance Layer

The AI-driven legal compliance layer is designed to enable the ERP system to ingest legislative updates, interpret regulatory implications, and generate compliance adaptation recommendations through explainable and auditable mechanisms. Operating independently from transactional ERP services, this layer supports continuous regulatory alignment while preserving institutional accountability [12].

Legislative data ingestion is managed through the Legal Data Interface, which connects to verified legal sources and applies NLP techniques to structure unstructured legal texts into machine-readable representations. Legal obligations and regulatory triggers are mapped into formal ontologies that encode relationships between legal entities, conditions, and operational processes [12]. In the proposed North Macedonian context, newly collected legal content would be compared with prior legislative versions to identify relevant changes [11].

The AI Reasoning and Adaptation Engine is designed to apply ontology-based reasoning, rule inference, and reinforcement learning to evaluate the impact of legal changes on ERP business processes and generate compliance adaptation recommendations. The proposed system would identify potentially impacted business domains and generate recommendations for process adaptation, such as revised payroll calculations or compliance-reporting procedures. Reviewer feedback and organizational data could be used in future implementation to refine interpretive accuracy over time [12].

To ensure legal reliability and ethical governance, all AI-generated recommendations are reviewed through the Human Oversight Layer before implementation. Legal experts or administrators can validate, reject, or modify proposed adaptations using interfaces that provide reasoning traces, legal references, and suggested system modifications. In the proposed workflow, approved changes would be logged through timestamped audit records containing legal and reviewer metadata, enabling traceability, compliance verification, and post-decision analysis [11-12].

Together, these mechanisms are intended to transform compliance management from a static configuration process into a continuous and context-aware capability connected to ERP operations.

C. Multilingual and Adaptive Components

The ERP framework incorporates multilingual and adaptive capabilities to support inclusive interaction and continuous legal alignment within North Macedonia's multilingual regulatory environment. The proposed framework includes real-time language switching as an intended multilingual capability, allowing users to interact with ERP services in their preferred language through AI-supported translation and interface localization mechanisms [11].

Multilingual NLP functionality enables the extraction and interpretation of legal obligations from legislative texts published in different official languages. These obligations are transformed into structured legal ontologies that support automated compliance reasoning and regulatory adaptation within ERP workflows [12].

Adaptive learning is conceptualized through the AI Reasoning and Adaptation Engine, which uses reinforcement learning and reviewer feedback to improve the accuracy of future compliance recommendations. In addition to legal adaptation, machine learning models could analyze organizational data related to employee roles, performance, and workload distribution to support predictive workforce management and operational optimization [11].

When legislative updates are detected, the proposed framework would compare new and previous legal versions, evaluate operational impacts, and generate adaptation recommendations such as revised payroll or leave-management policies. These recommendations are presented in the user's preferred language and validated by human reviewers prior to implementation [11-12].

By connecting multilingual processing and adaptive intelligence to the ERP architecture through an external compliance layer, the proposed framework is intended to support legally compliant, operationally responsive, and linguistically inclusive enterprise management.

The architecture described in this section represents the proposed design logic of the ERP compliance model. The current study evaluates this architecture through expert-based conceptual assessment rather than through deployment or operational prototype testing. Therefore, the described components should be interpreted as intended system capabilities requiring future implementation and empirical validation.

IV. Research Methodology

This section presents the methodology used to evaluate the proposed ERP model for adaptive legal compliance within the North Macedonian context. The study combines Design Science Research (DSR) principles with expert-based theoretical validation supported through a structured questionnaire. Because the research focuses on a conceptual ERP architecture integrating AI-driven legal reasoning, multilingual NLP, and adaptive learning, the evaluation emphasizes conceptual soundness, technical feasibility, and practical applicability rather than prototype implementation. The following sections describe the evaluation approach, expert selection process, questionnaire design, and data analysis procedures.

A. Evaluation Approach

The evaluation of the proposed ERP framework was grounded in Design Science Research (DSR), a methodology widely used in information systems research for the development and validation of innovative artifacts. The study focused on assessing a context-aware ERP architecture

integrating ontology-based legal reasoning, multilingual NLP, and adaptive learning mechanisms within enterprise compliance processes.

Because the research concerns a conceptual model rather than a deployed software implementation, expert-based validation was adopted as the primary evaluation strategy. Structured feedback was collected from professionals in ERP systems, artificial intelligence, legal compliance, software engineering, and public administration to assess the framework's conceptual integrity, feasibility, and practical relevance.

Data collection was conducted through a structured questionnaire containing both quantitative and qualitative components. This approach enabled systematic evaluation of the model's architectural coherence, implementation potential, and applicability within legally dynamic enterprise environments.

This evaluation approach is appropriate for assessing early-stage conceptual soundness and feasibility, but it does not replace prototype implementation, operational testing, or empirical validation in live ERP environments.

B. Expert Selection

Experts were selected using purposive sampling to ensure relevant expertise across ERP systems, software engineering, artificial intelligence, legal compliance, and public administration. Participants included ERP consultants, software engineers, IT architects, compliance specialists, legal experts, and academic professionals with experience in enterprise systems and regulatory processes.

All participants confirmed prior experience with ERP platforms, AI systems, legal compliance, or enterprise software environments. Respondents also reported their ERP-related experience levels, categorized as less than five years, five to ten years, or more than ten years, enabling consideration of professional experience during analysis.

Participation was conducted anonymously to encourage unbiased feedback and reduce social desirability bias. Respondents were informed that all collected data would be used exclusively for academic research purposes.

C. Questionnaire Design

The questionnaire was designed to evaluate the proposed ERP model across six dimensions: conceptual soundness, technical feasibility, AI-driven legal compliance, multilingual support, adaptability and scalability, and human oversight. The instrument combined closed-ended and open-ended questions to support both quantitative and qualitative evaluation.

The six questionnaire dimensions were derived from the main architectural and evaluation concerns of the proposed ERP compliance framework. Conceptual soundness assessed whether the model was understandable and internally coherent. Technical feasibility examined whether the architecture could plausibly be integrated into ERP environments. AI-driven legal compliance evaluated the perceived suitability of AI-assisted legal reasoning for compliance adaptation. Multilingual support reflected the framework's Macedonian–Albanian operating context. Adaptability and scalability assessed the model's ability to respond to changing legal and organizational conditions. Human oversight evaluated the accountability and validation mechanisms required for AI-assisted compliance decisions.

Closed-ended items used a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". These questions assessed the clarity of the model's architecture, the feasibility of integrating adaptive compliance mechanisms into ERP environments, the effectiveness of multilingual NLP support, and the role of human oversight in ensuring accountability and transparency within AI-assisted compliance processes.

The questionnaire also evaluated whether the proposed architecture could adapt to evolving legal and organizational conditions and whether its modular design supported scalable deployment across different institutional contexts. Two open-ended questions allowed participants to identify perceived strengths, limitations, and implementation challenges associated with the proposed framework.

This structure enabled a balanced evaluation of both the technical and contextual aspects of the ERP model while providing qualitative insights beyond numerical ratings.

The final dataset consisted of 51 questionnaire responses. The questionnaire included demographic and professional background items, 14 substantive closed-ended Likert-scale evaluation items, and two open-ended questions. All closed-ended items were complete for all 51 respondents. The first open-ended question received 34 valid responses, while the second received 31 valid responses. These response counts were considered during qualitative interpretation. Appendix A presents the questionnaire items used.

D. Instrument Validity and Reliability

Content validity was addressed by aligning the questionnaire dimensions with the main components and evaluation concerns of the proposed ERP compliance framework, including conceptual clarity, technical feasibility, AI-assisted legal compliance, multilingual support, adaptability and scalability, and human oversight. Internal consistency was assessed using Cronbach's Alpha across the 14 substantive Likert-scale items, producing a value of approximately 0.871, which indicates strong internal consistency for the current sample. However, construct validity was not independently verified through factor analysis, and the study did not include prototype testing or follow-up interviews. Therefore, the instrument should be interpreted as supporting expert-based conceptual evaluation rather than full psychometric or operational validation.

E. Data Collection and Analysis

Data was collected through a secure online questionnaire distributed via academic and professional networks. Participants received information regarding the study objectives, anonymity procedures, and research ethics prior to participation. The questionnaire remained open for two weeks, and no identifying information was stored. All closed-ended Likert-scale items were complete for all 51 responses and were included in the descriptive and reliability analysis. The dataset also contained a "Score" variable; however, because all entries contained a value of zero, this variable was excluded from analytical interpretation.

Quantitative responses were analyzed using descriptive statistics, including frequency distributions, mean scores, standard deviations, and 95% confidence intervals for each evaluation dimension. Dimension-level scores were calculated by aggregating thematically related Likert-scale items corresponding to the six evaluation dimensions. These measures enabled the identification of agreement levels, response patterns, and variability across expert assessments. Internal consistency of the questionnaire was evaluated using Cronbach's Alpha across the 14 substantive Likert-scale items. The resulting Cronbach's Alpha value was approximately 0.871, indicating strong internal consistency of the evaluation instrument.

Qualitative responses from the open-ended questions were analyzed using thematic coding to identify recurring strengths, limitations, and implementation concerns associated with the proposed ERP model. Common themes included the benefits of modular architecture and multilingual integration, as well as concerns related to AI interpretability, legal ontology maintenance, and integration complexity.

The combination of statistical and thematic analysis provided a multidimensional evaluation of the model's conceptual validity, technical feasibility, and contextual applicability.

V. Results

The evaluation assessed the proposed ERP model's conceptual validity, technical feasibility, and contextual suitability within the legal and organizational environment of North Macedonia. Expert feedback was analyzed through both quantitative ratings and qualitative reflections, providing insight into the model's architectural coherence, compliance functionality, multilingual capabilities, and implementation considerations. The findings highlight strong overall support for the proposed framework while also identifying challenges relevant to future implementation and validation efforts.

A. Participant Profile

The expert evaluation included participants from diverse professional backgrounds related to ERP systems, software engineering, artificial intelligence, legal compliance, and public administration. Software engineers represented the largest participant group, alongside ERP architects, lecturers, compliance specialists, and public-sector professionals. Additional expertise in AI engineering, cloud infrastructure, and network systems further broadened the evaluation perspective.

Among the 51 respondents, 26 reported less than five years of ERP experience, 17 reported between five and ten years of ERP-related experience, and 8 reported more than ten years of ERP experience. The participant structure, therefore, reflects perspectives from respondents with varying levels of ERP familiarity and implementation exposure.

Although 26 respondents reported fewer than five years of ERP-related experience, this distribution should be interpreted in light of the multidisciplinary nature of the evaluation. The proposed framework combines ERP architecture, software engineering, AI-assisted reasoning, legal compliance, multilingual support, and public-sector regulatory context. Therefore, participants with relevant experience in these related domains were considered suitable for assessing specific aspects of the model. Nevertheless, the relatively high proportion of respondents with limited ERP-specific experience is acknowledged as a limitation when interpreting conclusions about practical ERP deployment feasibility.

B. Quantitative Results

The quantitative evaluation demonstrated consistently positive assessments across all six dimensions of the proposed ERP model using a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". Mean scores exceeded 4.2 across all categories, indicating strong expert support for the model's conceptual structure, technical feasibility, and compliance-oriented functionality within the North Macedonian context.

The highest-rated dimension was Multilingual ERP Support with a mean score of 4.51, reflecting the perceived importance of Macedonian and Albanian language integration within enterprise systems. Participants identified multilingual NLP capabilities as essential for improving communication accuracy, operational inclusivity, and regulatory interpretation in bilingual environments.

Human oversight received the second-highest score at 4.49. Respondents emphasized the importance of maintaining expert validation for AI-generated compliance decisions through auditability, traceability, and accountability mechanisms. Conceptual Clarity followed with a mean score of 4.42, indicating strong agreement regarding the coherence and internal structure of the proposed architecture.

Technical feasibility achieved a mean score of 4.33, suggesting that the modular and service-oriented architecture could be integrated into existing ERP environments with appropriate organizational readiness and infrastructure support. Adaptability and Scalability received a mean score of 4.32, reflecting confidence in the framework's ability to respond to evolving legal and organizational requirements through adaptive learning mechanisms.

Finally, AI compliance validity received a mean score of 4.29, indicating positive assessments regarding the feasibility of AI-assisted legal interpretation and compliance automation. Participants nevertheless emphasized the continued necessity of combining AI reasoning with human oversight to preserve legal reliability and institutional trust.

The dimension-level results presented in Table I were constructed from thematically related questionnaire items. These dimensions include conceptual clarity, multilingual ERP support, technical feasibility, AI-assisted compliance adaptability, scalability, and human oversight mechanisms. The aggregation of questionnaire items into thematic dimensions was used to summarize expert evaluation trends across the proposed ERP framework components. The results are summarized in Table I.

In addition to the aggregated dimension scores, item-level descriptive statistics demonstrated consistently positive evaluations across the questionnaire items. The highest agreement was observed for the necessity of full Macedonian and Albanian language support, with a mean score of 4.57.

Strong agreement was also observed for multilingual NLP functionality (mean = 4.55) and the importance of human oversight in AI-assisted compliance adaptation (mean = 4.49). Comparatively lower, although still positive, agreement was observed for the feasibility of automated legislative interpretation (mean = 4.10), suggesting comparatively greater caution regarding autonomous legal interpretation mechanisms.

TABLE I. MEAN EXPERT EVALUATION SCORES ACROSS SIX ASSESSMENT DIMENSIONS OF THE PROPOSED ERP MODEL.

Evaluation Dimension	Mean score	SD	95% CI
Multilingual ERP support	4.51	0.59	4.34-4.67
Human oversight	4.49	0.76	4.28-4.70
Conceptual clarity	4.42	0.60	4.25-4.59
Technical feasibility	4.33	0.77	4.12-4.55
Adaptability and scalability	4.32	0.52	4.17-4.47
AI compliance validity	4.29	0.54	4.14-4.44

Note. N = 51; CI = confidence interval.

Table II presents selected item-level descriptive statistics for representative questionnaire items associated with multilingual support, human oversight, and AI-assisted compliance functionality. These item-level results complement the aggregated dimension-level findings by illustrating expert agreement patterns across specific aspects of the proposed ERP framework.

TABLE II. SELECTED ITEM-LEVEL DESCRIPTIVE STATISTICS

Questionnaire item	*Mean score*	*SD*	*95% CI*
Necessity of Macedonian and Albanian language support	4.57	0.83	4.33-4.80
Multilingual NLP functionality	4.55	0.61	4.38-4.72
Human oversight necessity	4.49	0.76	4.28-4.70
Automated legislative interpretation feasibility	4.10	0.76	3.89-4.31

Note. N = 51; CI = confidence interval.

C. Qualitative Feedback

The qualitative findings were derived from optional open-ended questionnaire responses. The first open-ended question received 34 valid responses, while the second received 31 valid responses. Consequently, the qualitative findings should be interpreted as supplementary thematic insights rather than complete-sample evidence.

The open-ended responses provided additional insight into the perceived strengths and limitations of the proposed ERP model. A dominant theme concerned the system's strong alignment with legal compliance requirements in environments characterized by frequent regulatory changes. Participants viewed the architecture as particularly relevant for organizations requiring adaptive compliance capabilities beyond those offered by traditional ERP systems.

Human oversight emerged as another central theme. Respondents consistently emphasized that AI-generated compliance recommendations should remain subject to expert validation in order to ensure accountability, legal reliability, and institutional trust. The inclusion of auditability and traceability mechanisms was therefore regarded as a major architectural strength.

Participants also highlighted the innovative integration of AI-driven legal reasoning, adaptive learning, and modular ERP services. The modular architecture was considered beneficial for scalability and integration with existing infrastructures, particularly within organizations operating with heterogeneous or legacy systems. Multilingual support for Macedonian and Albanian was similarly viewed as essential for operational inclusivity and accurate legal interpretation.

Despite the positive assessments, experts identified several implementation challenges. These included integration complexity, infrastructure requirements, AI interpretability limitations, and the risk of excessive reliance on automated legal reasoning. Respondents also noted that successful adoption would depend on organizational readiness, effective user training, and clear governance procedures. Data traceability and audit-ready compliance records were considered critical for maintaining long-term trust and regulatory accountability.

Overall, the qualitative feedback reinforced the quantitative findings while offering deeper insight into the practical conditions necessary for successful implementation of the proposed framework.

VI. DISCUSSION

The expert evaluation provides strong support for the proposed ERP model's conceptual structure, technical feasibility, and relevance within legally dynamic enterprise environments. The findings demonstrate broad agreement regarding the value of integrating AI-assisted compliance mechanisms, multilingual support, adaptive learning, and human oversight within ERP architectures. This section interprets these findings in relation to ERP system design, regulatory adaptability, and existing compliance approaches.

A. Interpretation of Expert Evaluation

The evaluation results provide consistent expert-based support for the proposed ERP framework across all assessed dimensions. Expert responses suggest that the model has the potential to address key challenges associated with regulatory adaptation, multilingual operation, and AI-assisted compliance management within enterprise systems. However, this support should be interpreted as conceptual validation rather than evidence of operational effectiveness in a deployed ERP environment.

Particular support was observed for the modular separation between operational ERP services and compliance-management functions. Participants viewed this architectural distinction as essential for enabling scalable and maintainable regulatory adaptation without disrupting transactional business processes. The integration of auditability, traceability, and human validation mechanisms further strengthened perceptions of legal reliability and institutional accountability.

The findings also suggest that the proposed framework aligns well with emerging expectations for intelligent ERP environments capable of supporting continuous compliance adaptation in complex regulatory settings. These findings

reinforce the model's relevance as a foundation for future implementation and empirical testing.

B. Implications for ERP Design

The evaluation highlights several broader implications for contemporary ERP system design. First, the findings suggest the potential feasibility of AI-assisted compliance management as an alternative to static and manually maintained regulatory mechanisms. The integration of ontology-based reasoning, adaptive learning, and multilingual NLP demonstrates how ERP systems can evolve toward continuous and context-aware legal alignment.

At the same time, the results strongly reinforce the necessity of human-in-the-loop governance within AI-driven enterprise environments. Experts consistently emphasized that automated legal interpretation must remain subject to human validation in order to preserve accountability, transparency, and institutional trust. Consequently, explainability, auditability, and reviewer oversight should be treated as foundational architectural requirements rather than optional enhancements.

The findings also underscore the importance of multilingual functionality in ERP systems operating within linguistically diverse regulatory contexts. In the North Macedonian setting, support for both Macedonian and Albanian was viewed as essential for inclusive interaction, accurate compliance interpretation, and organizational accessibility. More broadly, this suggests that multilingual NLP capabilities will become increasingly important in internationally distributed or multi-jurisdictional ERP deployments.

Overall, the study supports a transition from rigid ERP infrastructures toward adaptive enterprise ecosystems capable of combining operational efficiency with continuous regulatory responsiveness.

C. Comparison with Existing Approaches

Compared with traditional ERP systems that rely on static compliance rules and manually configured updates, the proposed framework introduces a more adaptive and context-aware approach to regulatory management. Conventional ERP architectures often struggle to accommodate evolving legal requirements without extensive redevelopment or operational disruption.

In contrast, the proposed model integrates AI-assisted legal reasoning, adaptive learning, multilingual NLP, and modular compliance services capable of responding dynamically to legislative changes. The separation between transactional ERP operations and compliance-management layers further improves scalability, maintainability, and deployment flexibility.

The architecture also addresses common limitations associated with opaque automation by incorporating auditability, traceability, and human oversight directly into the compliance workflow. These mechanisms are intended to enhance institutional accountability and distinguish the framework from conventional rule-based compliance systems that frequently lack transparent adaptation processes.

Collectively, these features position the proposed model as a promising conceptual contribution toward intelligent ERP systems capable of maintaining continuous legal alignment in multilingual and rapidly evolving regulatory environments.

VII. Threats to Validity

The expert-based evaluation of the proposed ERP model is subject to several validity considerations that should be acknowledged when interpreting the findings.

Regarding internal validity, the evaluation relied on expert judgment, meaning that responses were influenced by participants' professional experiences and familiarity with ERP systems, artificial intelligence, and compliance processes. Because the study assessed a conceptual framework rather than a deployed implementation, some variation in interpretation across respondents was expected.

A further limitation concerns the distribution of ERP-specific experience among participants. Although expertise was treated as multidimensional and included ERP systems, software engineering, artificial intelligence, legal compliance, public administration, and enterprise software experience, 26 of the 51 respondents reported fewer than five years of ERP-related experience. This provides useful multidisciplinary perspectives but may limit the strength of conclusions regarding senior ERP implementation feasibility.

Construct validity is also relevant due to the use of a 5-point Likert scale, where respondents may have interpreted agreement levels differently, particularly for abstract concepts such as adaptive learning or AI-driven legal compliance. Although the questionnaire was carefully structured around clearly defined evaluation dimensions, the absence of follow-up interviews may have limited clarification of individual interpretations.

External validity is constrained by the study's focus on the North Macedonian regulatory and linguistic environment, including bilingual support for Macedonian and Albanian. While core architectural principles such as modularity, AI-assisted compliance, and human oversight may be transferable to other contexts, the evaluation results remain closely connected to the regional conditions in which the framework was designed.

Several measures were applied to mitigate these limitations. Experts were selected based on relevant professional experience in ERP systems, information systems, or legal compliance, and participation was conducted anonymously to encourage unbiased responses. In addition, the combination of quantitative and qualitative feedback provided a broader and more consistent understanding of participant perspectives across different evaluation dimensions.

Although the analysis included mean scores, standard deviations, confidence intervals, and reliability assessment using Cronbach's Alpha, the study did not conduct factor analysis or subgroup comparisons by experience level. Therefore, the statistical findings should be interpreted as descriptive evidence of expert perceptions rather than as confirmatory psychometric validation or evidence of statistically significant differences between participant groups.

The study relies on descriptive and expert-based evaluation rather than inferential or experimental analysis. Consequently,

the dataset does not provide sufficient evidence for causal inference, predictive performance evaluation, or statistically verified implementation outcomes.

VIII. Conclusion and Future Work

This study presented an expert-based evaluation of a context-aware ERP model designed to support adaptive legal compliance within multilingual regulatory environments. The findings indicate strong expert-based support for the model's conceptual structure, technical feasibility, and contextual relevance within the North Macedonian enterprise landscape.

The proposed framework combines modular ERP architecture with AI-assisted legal reasoning, multilingual NLP, adaptive learning, and human oversight mechanisms and is intended to support continuous regulatory alignment. Experts particularly emphasized the importance of multilingual support and human validation in maintaining operational inclusivity, accountability, and trust within AI-assisted compliance processes.

The evaluation further suggests that adaptive and explainable compliance mechanisms represent an important direction for future ERP development, particularly in environments characterized by evolving legal requirements and multilingual governance structures. At the same time, the study identified implementation challenges related to integration complexity, organizational readiness, and AI interpretability, highlighting the need for careful governance and deployment planning.

Several limitations should be acknowledged. The evaluation focused on theoretical validation through expert assessment rather than empirical testing of a deployed system. Consequently, the findings reflect professional perceptions of feasibility and relevance rather than operational performance under real-world conditions.

Future work will focus on prototype implementation and empirical validation of the proposed framework in live enterprise environments. Additional research will examine the effectiveness of AI-driven compliance adaptation, multilingual NLP processing, and reinforcement-learning mechanisms over time. Cross-jurisdictional studies will also be necessary to evaluate the transferability of the architecture beyond North Macedonia and to identify adjustments required for broader regulatory and linguistic contexts.

Overall, the proposed model contributes toward the development of intelligent ERP systems capable of supporting continuous legal alignment, accountable automation, and multilingual enterprise management in dynamically evolving regulatory environments.

Appendix A: Questionnaire Items

The questionnaire used in the expert evaluation consisted of demographic/background questions, 14 closed-ended Likert-scale evaluation items, and two open-ended questions. The Likert-scale items were rated on a 5-point scale ranging from 'Strongly Disagree' to 'Strongly Agree'.

A. Background Information

Your professional role?

Years of experience with ERP systems?

I confirm that I have experience related to ERP systems, legal compliance, AI, or enterprise systems.

B. Likert-Scale Evaluation Items

The proposed ERP model is conceptually clear and logically structured.

The interaction between ERP modules and AI components is well-defined and understandable.

Automated interpretation of legislative changes is feasible within the proposed ERP model.

The AI-driven compliance component can improve legal alignment of ERP systems in North Macedonia.

Full support for both Macedonian and Albanian languages is necessary in ERP systems operating in North Macedonia.

NLP-based multilingual modules can improve clarity and reduce operational errors in ERP systems.

Adaptive learning mechanisms enable ERP systems to respond effectively to regulatory or organizational changes.

The proposed ERP model is scalable and suitable for enterprises in North Macedonia.

The proposed ERP model provides a coherent and realistic approach for integrating legal compliance, multilingual support, and adaptive learning within a single ERP system.

The proposed ERP model ensures that compliance-related decisions and process adaptations are traceable, auditable, and linked to specific legal requirements.

Human validation and oversight of AI-driven compliance adaptations are necessary and appropriate for ensuring accountability and trust in ERP systems.

The proposed ERP model appears technically feasible to integrate with existing ERP systems using modular services and standard interfaces.

Multilingual NLP support can improve data consistency and accuracy across Macedonian and Albanian records and reports.

The proposed ERP model sufficiently mitigates risks related to AI-based legal misinterpretation through human validation and auditability.

C. Open-Ended Questions

What do you consider the main strength or limitation of the proposed ERP model?

Which aspect would require the most attention in a real-world implementation of this model?

References

[1] A. S. Sayied, F. Batool, and R. M. Butt, "Business Process Automation: Compliance Management," Journal of Business and Management Research, vol. 4, no. 2, pp. 801–836, 2025, doi: 10.64105/jbmr.04.02.443.

[2] E. Bajrami and F. Halili, "Exploring the Impact of Artificial Intelligence on the IoT and Digital Agenda in the Western Balkans: Integrating A Proposed Web Application for Regional Advancement," Journal of Natural Sciences and Mathematics of UT-JNSM, vol. 9, no. 17–18, pp. 244–258, 2024, doi: 10.62792/ut.jnsm.v9.i17-18.p2819.

[3] A. Beheshti, J. Yang, Q. Z. Sheng, B. Benatallah, F. Casati, S. Dustdar, H. R. M. Nezhad, X. Zhang, and S. Xue, "ProcessGPT: Transforming Business Process Management with Generative Artificial Intelligence," arXiv, 2023, doi: 10.48550/arXiv.2306.01771.

[4] V. Bezditnyi, "Use of Artificial Intelligence for Tax Planning Optimization and Regulatory Compliance," Research Corridor Journal of Engineering Science, vol. 1, no. 1, Jan. 2024, doi: 10.66320/8qe5f566.

[5] J. Bharadiya, "The Impact of Artificial Intelligence on Business Processes," European Journal of Technology, vol. 7, no. 2, pp. 15–25, 2023, doi: 10.47672/ejt.1488.

[6] M. Dumas, F. Fournier, L. Limonad, A. Marrella, M. Montali, J.-R. Rehse, R. Accorsi, D. Calvanese, G. De Giacomo, D. Fahland, A. Gal, M. La Rosa, H. Völzer, and I. Weber, "AI-Augmented Business Process Management Systems: A Research Manifesto," 2022, doi: 10.48550/ARXIV.2201.12855.

[7] E. Fabbri, L. Živković, D. Štrbac, and I. Ljumović, Regional potential for ICT collaboration in the context of Smart Specialisation in the Western Balkans. Publications Office of the European Union, 2025. [Online]. Available: http://ebooks.ien.bg.ac.rs/2205/

[8] P. Hehnle and M. Reichert, "Flexible process variant binding in information systems with software product line engineering," Journal of Systems and Software, vol. 230, p. 112530, 2025, doi: 10.1016/j.jss.2025.112530.

[9] M. Kastrati and M. Biba, "Natural language processing for Albanian: A state-of-the-art survey," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 6, p. 6432, 2022, doi: 10.11591/ijece.v12i6.pp6432-6439.

[10] U. A. Marovac, A. R. Avdić, and N. Lj. Milošević, "A Survey of Resources and Methods for Natural Language Processing of Serbian Language," arXiv, Version 1, 2023, doi: 10.48550/ARXIV.2304.05468.

[11] S. Miftari and A. Aliu, "Designing a Tailored ERP System with Localized Architecture for North Macedonian Enterprises," in 2025 MIPRO 48th ICT and Electronics Convention, 2025, pp. 1991–1996, doi: 10.1109/MIPRO65660.2025.11131743.

[12] S. Miftari, A. Aliu, and A. Luma, "A context-aware framework for modeling ERP business process adaptation under national legal requirements," International Journal of Interactive Mobile Technologies (iJIM), vol. 20, no. 13, May 2026. [Online]. Available: https://online-journals.org/index.php/i-jim/article/view/62041.

[13] N. Milinković and M. Galjak, "A case study on AI usage for collecting philanthropy data in the Western Balkans," in The Routledge Handbook of Artificial Intelligence and Philanthropy, 1st ed., G. Ugazio and M. Maricic, Eds. Routledge, 2024, pp. 240–256, doi: 10.4324/9781003468615-18.

[14] H. P. Kothandapani, "AI-Driven Regulatory Compliance: Transforming Financial Oversight through Large Language Models and Automation," Emerging Science Research, vol. 3, Jan. 2025, Art. no. 12. doi: 10.6084/M9.FIGSHARE.28251608.

[15] M. Rosemann, J. V. Brocke, A. Van Looy, and F. Santoro, "Business process management in the age of AI – three essential drifts," Information Systems and E-Business Management, vol. 22, no. 3, pp. 415–429, 2024, doi: 10.1007/s10257-024-00689-9.

[16] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, "A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions," IEEE Open Journal of the Computer Society, vol. 4, pp. 280–302, 2023, doi: 10.1109/OJCS.2023.3300321.

[17] H. Aşan and V. Tecim, "Model Design and Development for a User Support System Using Artificial Intelligence Techniques in Enterprise Resource Planning Software," International Journal of Contemporary Economics and Administrative Sciences, vol. 13, no. 2, pp. 836–864, Jan. 2024, doi: 10.5281/zenodo.10476230.

AI-Driven Service Innovation, Customer Satisfaction, and Guest Loyalty: Evidence from Shenyang Airport Hotel, China

Dayong Zu[1], Kawalin Angkananon[2]*, Yoksamon Jeaheng[3]
International School of Tourism, Suratthani Rajabhat University, Surat Thani, Thailand[1, 3]
Faculty of Management Sciences, Suratthani Rajabhat University, Surat Thani, Thailand[2]

***Abstract*—This study investigated how guests' technological perceptions of Artificial Intelligence (AI) applications perceived usefulness (PU), perceived ease of use (PEOU), enjoyment (ENJ), and privacy concerns (PC) influence customer satisfaction (CS) and customer loyalty (CL) at the Shenyang Airport Hotel, a three-star airport property in a secondary Chinese city. Grounded in an integrated framework combining the Technology Acceptance Model (TAM) with Customer Experience Theory. A quantitative survey of 452 guests who had experienced AI-enabled services during their stay was conducted. Data were analysed using reliability assessment, exploratory factor analysis (to verify the factor structure in a novel context), Pearson correlation, multiple regression, and bootstrap mediation analysis. The results indicated that perceived usefulness ($\beta = 0.314$, $p < 0.001$), enjoyment ($\beta = 0.141$, $p < 0.01$), and perceived ease of use ($\beta = 0.127$, $p < 0.01$) each positively influenced customer satisfaction, whereas privacy concerns exerted a significant negative effect ($\beta = -0.205$, $p < 0.001$). Customer satisfaction, in turn, significantly predicted customer loyalty ($\beta = 0.409$, $p < 0.001$), and partially mediated all four perception-to-loyalty pathways. One-way ANOVA further revealed significant differences in customer loyalty across age groups, educational levels, and income brackets. Theoretically, this study extended TAM by simultaneously incorporating affective and security-related dimensions into AI service acceptance within an underexamined hospitality context. Practically, the findings offer airport hotel managers in secondary cities evidence-based guidance on prioritising utility-enhancing AI features, simplifying service interfaces, enriching hedonic engagement, and communicating data governance policies transparently to mitigate privacy-driven dissatisfaction. The study was limited by its single-site, cross-sectional design; future research should adopt longitudinal or multi-site approaches to strengthen generalisability.**

Keywords*—*Artificial intelligence; airport hotels; technology acceptance model; customer satisfaction; customer loyalty; secondary city; China

I. Introduction

AI has rapidly become a transformative force in the global hospitality industry. From AI-powered chatbots and facial recognition check-in systems to service robots and predictive analytics platforms, hotels are deploying AI across the full service cycle to enhance operational efficiency, personalize guest experiences, and strengthen competitive positioning [1-2]. In China, this transformation has been further accelerated by strong government policy support and advanced digital infrastructure. Luxury flagship hotels in first-tier cities exemplify the frontier of this trend: Alibaba's FlyZoo Hotel in Hangzhou offers fully AI-integrated experiences spanning autonomous check-in, voice-controlled rooms, and robot concierge services [3], while international chains such as Marriott, Hilton, and Inter Continental have deployed AI-driven dynamic pricing and sentiment analysis platforms across their premier Chinese properties [4]. These innovations have generated a corresponding surge of scholarly interest, with empirical studies consistently reporting that well-implemented AI services improve guest satisfaction and stimulate revisit intention [3, 5].

Despite this growing body of practice and research, the existing literature reflects a pronounced geographical bias. The overwhelming majority of empirical studies on AI in Chinese hotels have concentrated on luxury or upscale establishments in major metropolitan centers such as Shanghai, Beijing, and Guangzhou, where hotels operate with sophisticated technological infrastructure and technologically fluent guest bases [5]. Airport hotels in secondary Chinese cities have received considerably less academic attention. Shenyang, a major industrial and transportation hub in northeastern China, exemplifies such a context. The Shenyang Airport Hotel, located adjacent to Shenyang Taoxian International Airport, operates as a typical three-star airport hotel that serves a mix of business travelers, transit passengers, and increasingly leisure tourists. Despite its strategic location and stable customer flow, the hotel faces persistent operational challenges. According to the hotel's official operational records and a preliminary field survey conducted by the authors in March 2026, which included structured interviews with the hotel manager and three senior front-desk staff members, the hotel's annual repurchase rate stood at only 18% in 2024, substantially below the national average for comparable domestic airport hotels [6-7]. The same sources further reveal the absence of a systematic Customer Relationship Management (CRM) system or loyalty program, and limited technological infrastructure capable of supporting personalized guest retention strategies. These operational realities create both a practical imperative and a research opportunity: to understand how guests perceive AI applications in such a setting and to examine whether these perceptions translate into higher satisfaction and greater loyalty.

There are three research gaps addressed in this study. First, while prior studies have confirmed that AI can improve service efficiency and personalization [2, 5], few have examined how individual technological perceptions, specifically whether an AI

system is perceived as useful, easy to use, enjoyable, or privacy-invasive, affect satisfaction and loyalty outcomes in airport hotel contexts. Second, the majority of empirical work has concentrated on first-tier cities or upscale establishments, leaving the dynamics of AI perception formation in secondary-city, mid-scale airport hotels largely unexplored. Third, the mediating role of customer satisfaction in translating technological perceptions into loyalty intentions, although theoretically plausible, has not been rigorously tested in this particular context. Consequently, hotel managers in secondary cities often lack evidence-based guidance on which AI features to prioritize and how to address guests' privacy concerns when deploying AI services.

To fill these gaps, this study empirically investigates how four dimensions of guests' technological perceptions, perceived usefulness, perceived ease of use, enjoyment, and privacy concerns, influence customer satisfaction and, in turn, customer loyalty at Shenyang Airport Hotel. A quantitative survey of 452 guests who had used AI-enabled services during their stay was conducted between March and May 2026. The study addresses six research questions: 1) To what extent does the perceived usefulness of AI applications influence customer satisfaction among hotel guests? 2) To what extent does the perceived ease of use of AI applications influence customer satisfaction among hotel guests? 3) To what extent does the enjoyment derived from AI applications influence customer satisfaction among hotel guests? 4) To what extent do privacy concerns regarding AI applications negatively influence customer satisfaction among hotel guests? 5) Does customer satisfaction positively influence guests' behavioral intentions? 6) Does customer satisfaction mediate the relationships between guests' technological perceptions of AI applications, namely perceived usefulness, perceived ease of use, enjoyment, and privacy concerns, and customer loyalty?

The findings contribute to the literature in three ways. Theoretically, this study extends existing technology acceptance research by incorporating both affective (enjoyment) and security-related (privacy concern) factors into a single model applied to an AI service context that has been largely absent from prior scholarship. Empirically, it provides the first systematic investigation of AI-driven satisfaction and loyalty dynamics in a secondary-city Chinese airport hotel, a context characterized by partial AI infrastructure and low guest retention rates. Practically, it offers hotel managers in analogous settings concrete, evidence-based guidance on AI service design, privacy communication, and demographic differentiated implementation strategies. This study is organized as follows. Section II reviews the relevant literature and develops the study hypotheses. Section III describes the research methodology. Section IV presents the empirical results. Section V discusses the key findings and their theoretical and managerial implications. Section VI presents conclusions. Section VII outlines the research implementation, and Section VIII addresses the limitations of the study and suggests directions for future research.

II. Literature Review

A. AI Adoption in Hospitality and the Airport Hotel Context

AI integration in hotel operations has advanced from rule-based chatbot automation to adaptive systems spanning natural language processing, predictive analytics, and robotics [2, 4]. It falls into three categories: guest interaction systems (chatbots, facial recognition check-in), analytical platforms (revenue management, sentiment analysis), and automated physical operations (service robots, smart room control). Together, these shift service delivery from human-to-human to human-to-machine interaction, introducing efficiency gains alongside new psychological challenges for guests [1]. Airport hotels occupy a structurally distinctive niche, characterized by high turnover, short stays, and guests who prioritize efficiency over experience [8]. This makes them receptive to time-saving AI tools; yet the technological infrastructure of secondary-city airport hotels frequently lags behind first-tier counterparts, creating a context where AI effectiveness is uncertain, and guest perceptions may differ markedly from those documented in luxury urban hotels [5]. Table I contextualizes the AI service profile of the Shenyang Airport Hotel (SAH) against ten landmark hospitality studies. The three features operational at SAH: AI chatbots, facial recognition check-in, and automated check-in/check-out—are among the most frequently documented (frequency 4–5), confirming the study's empirical relevance. The absence of post-stay follow-up systems, recommendation engines, and AI-enabled loyalty programmes delineates the current technological gap and reinforces the study's practical motivation [6-7].

TABLE I. AI Service Features in the Hospitality

No	AI Feature / Application	[4]	[9]	[10]	[11]	[12]	[13]	14]	[15]	[16]	[17]	Authors
1	AI Chatbots	✓	✓		✓	✓	✓					✓
2	Post-stay Follow-up AI		✓			✓			✓	✓		
3	Facial Recognition Check-in	✓		✓		✓		✓				✓
4	Service Robots	✓	✓	✓							✓	
5	AI Recommendation Engines		✓	✓	✓		✓		✓			
6	Automated Check-in/Check-out		✓			✓		✓		✓		✓
7	AI Concierge Systems			✓		✓	✓				✓	
8	Biometric Room Access					✓		✓			✓	
9	Hyper-personalized Services		✓		✓		✓		✓			
10	AI-enabled Loyalty Programme							✓	✓		✓	

B. TAM and Its Extensions

Davis [26] established that PU and PEOU jointly determine technology acceptance. Subsequent extensions accommodate the hedonic and risk dimensions of consumer service contexts [18]. Two extensions are central here. First, perceived enjoyment (ENJ), intrinsic positive affect generated by the AI interaction, independently enhances acceptance in service settings where novelty and interactivity are salient. Second, privacy concern, unease about biometric data collection and use, functions as a consistent negative moderator, particularly acute in airport environments where facial recognition is operationally embedded [19, 21]. Integrating all four perceptions into a single model represents the primary theoretical contribution of this study.

C. Customer Satisfaction and Customer Loyalty

Customer satisfaction is conceptualized as a post-consumption evaluative judgment reflecting whether AI-mediated service encounters have met or exceeded guest expectations [20]. Customer Experience Theory [22] situates this judgment within a multi-dimensional architecture—cognitive, emotional, social, and physical, explaining divergent evaluations of the same AI features across guest segments. In the present study, CET serves as an interpretive scaffold rather than an independently operationalized measurement framework: it motivates the inclusion of both utilitarian (PU, PEOU) and hedonic (ENJ) dimensions alongside a risk-based barrier (PC), and it explains why the same AI features may generate divergent satisfaction evaluations across guest segments defined by age, education, and income. The four TAM-derived constructs are the operationalized variables; CET provides the theoretical rationale for their simultaneous inclusion and for the expectation that satisfaction will mediate their effects on loyalty. Future research could operationalize CET dimensions directly, for example, through validated cognitive and emotional appraisal sub-scales to permit a more complete test of the integrated framework. Customer loyalty is operationalized as revisit intention and word-of-mouth (WOM) recommendation, capturing both the direct economic value and the social amplification of loyalty [23]. Whether technological perceptions shape loyalty primarily through satisfaction or also through direct paths is addressed via bootstrap mediation analysis.

TABLE II. LITERATURE REVIEW AND RESEARCH GAPS

Theme / Construct	Key Finding from Literature	[12]	[8]	[5]	[19]	[21]	Authors
Perceived Usefulness (PU)	Strongest positive predictor of AI-service satisfaction; time-saving and accuracy are most valued.	✓		✓			✓
Perceived Ease of Use (PEOU)	Positively influences satisfaction; interface complexity deters older and less tech-savvy guests.	✓		✓			✓
Enjoyment / Hedonic Motivation	Independently enhances satisfaction beyond utilitarian perceptions; salient in service-intensive encounters.		✓				✓
Privacy Concern	Negatively moderates AI acceptance; biometric data collection intensifies unease without transparent policies.				✓	✓	✓
Satisfaction → Loyalty Mediation	Satisfaction partially mediates the perception-to-loyalty path; direct effects also exist but are weaker.			✓			✓
Context: First-tier / Upscale Hotel	Empirical base dominated by 5-star or chain hotels in major metropolitan centers.	✓	✓	✓	✓	✓	
Context: Airport Hotel, Secondary City	Partial AI infrastructure; high turnover; time-sensitive guests; low CRM maturity — underexamined.						✓
Demographic Moderation (ANOVA)	Age, education, and income moderate technology acceptance and loyalty levels.					✓	✓

Table II maps the thematic coverage of prior literature against the present study, revealing three interrelated gaps: 1) no prior study has examined PU, PEOU, ENJ, and PC simultaneously within an airport hotel model; 2) all comparable studies were conducted in first-tier cities or upscale segments, leaving secondary-city airport hotels unexamined; and 3) the mediating role of satisfaction has not been rigorously tested with bootstrap procedures in this context. Six hypotheses follow directly.

III. RESEARCH METHODOLOGY

This section describes the research design, study context, population, and sampling strategy, instrument development, data collection, and statistical analysis procedures employed in the study.

A. Research Design

A quantitative, cross-sectional survey design was adopted to test the hypothesized relationships among guests' AI service perceptions, customer satisfaction, and customer loyalty. Quantitative methods are well-suited to evaluating predictive relationships among multiple continuous constructs and to drawing population-level inferences from probability-based samples [24]. Primary data were collected through a structured, self-administered questionnaire available in both digital and paper formats.

B. Research Context: Shenyang Airport Hotel

Shenyang Airport Hotel is a three-star property located approximately one kilometer from Terminal 3 of Shenyang Taoxian International Airport, Liaoning Province. The hotel operates 163 rooms, of which 123 are standard, 30 are business, and 10 are family, and all are equipped with smart door locks and smart televisions. The property operates without dedicated IT personnel, a CRM system, or a formal loyalty programme. Its 2024 annual repurchase rate of 18% falls substantially below the national benchmark for comparable airport hotels. A preliminary field survey and management interviews conducted in March 2026 confirmed that three AI services are currently operational: an AI chatbot, a facial recognition self-check-in terminal, and an automated check-in/check-out kiosk—while more advanced retention-oriented AI tools remain absent [6-7].

C. Research Participants: Population and Sample Size

The target population comprised guests who had stayed at the hotel at least once within the two years preceding data collection and had used at least one AI-enabled service during their stay. Because the eligible population is continuous and uncountable, it was treated as theoretically infinite. The minimum sample size of 384 was determined using Cochran's formula [25].

To compensate for anticipated incomplete responses, 500 questionnaires were distributed. After data screening, 452 valid responses were retained (valid response rate = 90.4%), exceeding both the Cochran minimum (384) and the recommended threshold for the planned regression analyses.

D. Sample Size Sampling Procedure

Stratified random sampling was applied to ensure proportionate representation of key guest subgroups. The sampling frame was stratified along three dimensions: travel purpose, age group, and nationality. Within each stratum, respondents were selected randomly. This approach reduces the risk of segment-driven bias and strengthens the external validity of the findings [25].

E. Research Instrument Development and Validation

1) Item generation: All measurement items were adapted from validated scales: PU and PEOU from Davis [26]; ENJ from [18]; PC from [19]; CS from [20]; and CL from Zeithaml et al. [23]. Items were translated into Simplified Chinese and independently back-translated by a second bilingual researcher to ensure semantic equivalence [27].

2) Expert validation: Three subject-matter experts assessed content validity using the Index of Item-Objective Congruence (IOC). The mean IOC across all items was 0.82, exceeding the 0.70 threshold [28]. Two items with IOC values below 0.70 were removed, yielding a final questionnaire of 40 items organized into four sections: 1) sociodemographic and travel characteristics; 2) AI usage history; 3) six Likert-scale constructs rated on a five-point scale (1 = strongly disagree to 5 = strongly agree); and 4) open-ended qualitative questions.

3) Reliability and validity test: Cronbach's alpha coefficients for all multi-item constructs exceeded .85, demonstrating excellent internal consistency. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .895, and Bartlett's test of sphericity was statistically significant (χ^2 = 7283.729, df = 253, $p < .001$), confirming factorability of the correlation matrix. Exploratory factor analysis (EFA) using principal component extraction with varimax rotation extracted six factors with eigenvalues greater than 1, collectively accounting for 78.16% of the total variance. All items loaded substantially on their hypothesized factors ($\lambda > .70$), with negligible cross-loadings, supporting both convergent and discriminant validity [29]. EFA was selected over confirmatory factor analysis (CFA) because this study, although guided by theoretically motivated construct definitions, employs items adapted from multiple prior scales and tested together for the first time in the specific context of a secondary-city Chinese airport hotel. EFA was therefore appropriate to verify the empirical factor structure and item behaviour in this new setting before structural relationships were estimated [29]. Future research replicating the model in comparable contexts should employ CFA or structural equation modelling (SEM) to confirm construct dimensionality and provide stronger evidence of model fit.

F. Research Hypothesis

- H1: Perceived usefulness of AI applications has a positive effect on customer satisfaction.
- H2: Perceived ease of use of AI applications has a positive effect on customer satisfaction.
- H3: Enjoyment of AI applications has a positive effect on customer satisfaction.
- H4: Privacy concerns regarding AI applications harm customer satisfaction.
- H5: Customer satisfaction positively affects customer loyalty.
- H6: Customer satisfaction mediates the relationship between guests' technological perceptions of AI applications (perceived usefulness, perceived ease of use, enjoyment, privacy concerns) and customer loyalty.

These hypotheses are illustrated in the conceptual model below (Fig. 1).

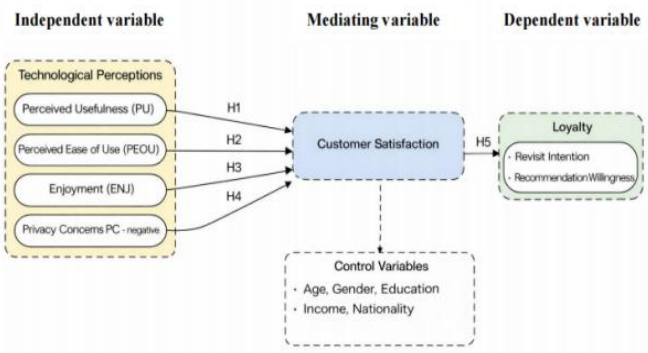


Fig. 1. Research framework of AI perceptions, customer satisfaction, and loyalty.

G. Data Collection

Data were collected between March and May 2026 through four complementary channels: 1) email invitations to opted-in prior guests; 2) QR-code cards displayed at the front desk, lobby, and in each guest room; 3) a pinned post on the hotel's official WeChat account; and 4) paper questionnaires available at the front desk for guests who preferred offline completion. Respondents received a 15% discount voucher for a future stay as an incentive. Ethics approval was obtained from the Research Ethics Committee of Suratthani Rajabhat University (Approval No.: SRU-EC 2026/029). All participation was voluntary and anonymous.

H. Data Analysis

All statistical analyses were performed using Statistics software, following a sequential six-stage procedure:

1) Descriptive statistics: Means, standard deviations, and frequency distributions were computed for all constructs and demographic variables.

2) Reliability and validity: Cronbach's alpha was recomputed for the full sample. Exploratory Factor Analysis (EFA) with principal component extraction and varimax rotation was conducted; the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity verified sampling adequacy [30].

3) Correlation analysis: Pearson's r coefficients were computed for all construct pairs; Variance Inflation Factor (VIF) values confirmed the absence of multicollinearity (threshold VIF < 10) [29].

4) Hypothesis testing (H1–H5): Multiple linear regression examined the effects of PU, PEOU, ENJ, and PC on CS (H1–H4). Simple linear regression examined the effect of CS on CL (H5a–H5b). Standardized beta coefficients (β) and their significance levels were reported.

5) Mediation analysis (H6): Bootstrap mediation with 5,000 resamples was conducted following Preacher and Hayes [31]. Bias-corrected 95% confidence intervals (CIs) were computed; an indirect effect was deemed significant if the CI excluded zero. Partial mediation was inferred when both the direct and indirect effects remained significant.

6) Difference analysis: One-way ANOVA examined whether CL differed significantly across age groups, educational levels, and income brackets. Where F-ratios were significant, Scheffé post-hoc tests identified specific group differences.

All p-values are two-tailed; the significance level was set at $\alpha = 0.05$ throughout the study.

IV. Results

A. Participant Characteristics

This study examined the effects of AI adoption in hotel management on guests' perceptions, satisfaction, and loyalty. Data were collected from guests who had experienced AI-enabled services at Shenyang Airport Hotel. The majority of respondents were male (52.65%), aged between 30 and 49 years (53.10%), and held at least a bachelor's degree (86.94%). Business travel was the predominant purpose of stay (42.48%), and 69.03% of respondents had stayed at the hotel at least once within the preceding two years. Detailed demographic information is presented in Table III.

Table III shows that the majority stayed at the hotel once per year (44.36%), while 48.04% reported using AI services on more than two occasions during their stay. The most commonly encountered AI technology was the facial recognition self-check-in/check-out terminal (68.63%). The primary channels through which respondents became aware of the hotel's AI services were staff recommendations at check-in (68.14%) and prompts displayed on the self-check-in terminal (61.52%). These findings indicate that the sample possessed a sufficient level of familiarity with AI adoption in hotel management to support the validity of subsequent empirical analyses.

TABLE III. Demographic Profile of Respondents

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	238	52.65
	Female	214	47.35
Age	18–29 years	119	26.33
	30–49 years	240	53.10
	50–60 years	93	20.58
Education	High school or below	59	13.06
	Bachelor's degree	220	48.67
	Master's degree or above	173	38.27
Travel purpose	Business	192	42.48
	Leisure	156	34.51
	Transit	71	15.71
	Other	33	7.30
Prior hotel stays (past 2 years)	At least once	312	69.03
	More than a year	140	30.97

B. Descriptive Statistics, Reliability, and Validity

Table IV shows the mean scores of all six constructs. Customer Satisfaction recorded the highest mean ($\bar{x} = 3.95$), followed by Enjoyment ($\bar{x} = 3.88$), Perceived Usefulness ($\bar{x} = 3.87$), Perceived Ease of Use ($\bar{x} = 3.83$), and Customer Loyalty ($\bar{x} = 3.76$), respectively. Privacy Concerns yielded the lowest mean ($\bar{x} = 2.98$), the only construct falling below the scale midpoint of 3.00.

TABLE IV. Descriptive Statistics and Reliability Coefficients

Variable	$\bar{x}$	SD	α	No. of Items
PU	3.87	0.70	0.859	3
PEOU	3.83	0.85	0.906	3
ENJ	3.88	0.83	0.904	3
PC	2.98	1.09	0.900	4
CS	3.95	0.61	0.891	5
CL	3.76	0.77	0.927	5

Table V shows item-level mean scores across all six constructs. Within Perceived Usefulness, time-saving was rated most favourably, affirming speed as the primary utilitarian benefit. For Perceived Ease of Use, unassisted terminal operation recorded the highest item mean, indicating that most guests navigated the interface independently. Novelty dominated the Enjoyment dimension, reflecting the affective appeal of facial recognition technology. All Privacy Concerns items fell around the scale midpoint, confirming a lack of unease over biometric data governance. Terminal accuracy drove Customer Satisfaction most strongly, while willingness to revisit yielded the highest item mean within Customer Loyalty, though preference over competitors remained comparatively low, suggesting limited competitive differentiation.

TABLE V. DESCRIPTIVE STATISTICS OF EACH ASSESSMENT ITEM

Assessment Items	$\bar{x}$	SD
Perceived Usefulness (PU)	**3.87**	**0.70**
1. The facial recognition self-check-in/check-out terminal saved me time compared to manual front desk service.	3.93	0.79
2. Using the AI terminal improved the efficiency of my check-in/check-out process.	3.88	0.64
3. The AI terminal met my need for quick service (e.g., when I was in a hurry for a flight).	3.81	0.67
Perceived Ease of Use (PEOU)	**3.83**	**0.85**
1. I could operate the facial recognition terminal without staff assistance.	3.86	0.96
2. The terminal's instructions (e.g., interface prompts for ID scanning) were clear and easy to understand.	3.84	0.80
3. Even if I had limited experience with technology, I could still operate the AI terminal easily.	3.79	0.79
Enjoyment (ENJ)	**3.88**	**0.83**
1. Using the facial recognition terminal added a sense of novelty to my stay.	3.98	0.92
2. I felt relaxed using the terminal (e.g., no need to wait in line for the front desk).	3.80	0.80
3. The AI terminal made me perceive the hotel as providing novel services.	3.87	0.77
Privacy Concerns (PC)	**2.98**	**1.09**
1. I worried about the hotel misusing my facial recognition data.	2.93	1.23
2. I did not fully trust that the hotel's AI system would protect my personal data.	2.96	1.01
3. The hotel did not clearly inform me about my facial data storage and deletion policy.	2.94	1.04
4. I am not concerned about how my facial data is stored.	3.07	1.09
Customer Satisfaction (CS)	**3.95**	**0.61**
1. I was satisfied with the accuracy of the facial recognition terminal (e.g., first-try verification).	4.00	0.71
2. I was satisfied with the AI terminal's response speed (e.g., fast room key issuance).	3.93	0.56
3. Using the AI terminal enhanced my overall stay experience.	3.93	0.55
4. Even without using AI services, I was satisfied with the hotel's general services (e.g., airport shuttle).	3.92	0.61
5. Overall, I was highly satisfied with my stay at Shenyang Airport Hotel.	3.96	0.63
Customer Loyalty (CL)	**3.76**	**0.77**
1. I am willing to stay at this hotel again on my next visit.	3.84	0.91
2. I will recommend this hotel to friends/family needing nearby accommodation.	3.73	0.73
3. I will prioritize this hotel for future Shenyang trips (business/transit).	3.74	0.71
4. I am willing to share positive feedback (including AI services) on social media.	3.76	0.74
5. The hotel's services make me more likely to choose it over nearby competitors.	3.72	0.75

C. Correlation Analysis

Table VI shows the Pearson correlation matrix for all key variables. Perceived usefulness, perceived ease of use, and enjoyment were each positively and significantly correlated with customer satisfaction ($r = 0.439$, 0.320, and 0.330, respectively; all $p < 0.01$), whereas privacy concerns exhibited a significant negative correlation with satisfaction ($r = -0.299$, $p < 0.01$). Customer satisfaction and customer loyalty demonstrated a moderate positive association ($r = 0.409$, $p < 0.01$). All inter-variable correlation coefficients fell below 0.80, indicating that multicollinearity was not a concern in the subsequent regression analyses.

TABLE VI. PEARSON CORRELATION MATRIX OF KEY VARIABLES

Variable	PU	PEOU	ENJ	PC	CS
PU	1				
PEOU	.313**	1			
ENJ	.316**	.507**	1		
PC	−.199**	−.112*	−.124**	1	
CS	.439**	.320**	.330**	−.299**	1
CL	.386**	.336**	.301**	−.327**	.409**

[a] Note: **p < 0.01, *p < 0.05.

D. Hypothesis Testing: H1–H4

Multiple regression analysis was conducted with customer satisfaction as the dependent variable and the four technological perception dimensions as predictors. The overall model was statistically significant, $F(4, 447) = 44.836$, $p < 0.001$, accounting for 28.6% of the variance in customer satisfaction ($R^2 = 0.286$). The regression results are summarised in Table VII.

TABLE VII. MULTIPLE REGRESSION RESULTS: AI PERCEPTIONS PREDICTING CUSTOMER SATISFACTION (H1–H4)

Predictor	B	SE	β	t	p
Constant	2.475	0.192	—	12.889	< 0.001
PU (H1)	0.274	0.038	0.314	7.222	< 0.001
PEOU (H2)	0.092	0.034	0.127	2.682	0.008
ENJ (H3)	0.105	0.035	0.141	2.991	0.003
PC (H4)	−0.116	0.023	−0.205	−5.017	< 0.001

Perceived usefulness exerted the strongest positive effect on customer satisfaction ($\beta = 0.314$, $p < 0.001$), followed by enjoyment ($\beta = 0.141$, $p < 0.01$) and perceived ease of use ($\beta = 0.127$, $p < 0.01$). Privacy concerns had a significant negative effect on satisfaction ($\beta = -0.205$, $p < 0.001$). Accordingly, Hypotheses H1, H2, H3, and H4 were all supported.

E. Hypothesis Testing: H5

Simple regression analysis revealed that customer satisfaction exerted a significant positive effect on customer loyalty ($\beta = 0.409$, $t = 9.501$, $p < 0.001$). The model explained 16.7% of the variance in customer loyalty ($R^2 = 0.167$, $F = 90.278$, $p < 0.001$). These findings support Hypothesis H5, confirming that guests who reported higher satisfaction with AI-enabled hotel management services were more likely to express intentions to revisit the hotel and to recommend it to others through word-of-mouth.

F. Hypothesis Testing: H6

To test Hypothesis H6, a bootstrap mediation analysis was performed using 5,000 resamples, following the procedure recommended by Preacher and Hayes. Bias-corrected 95% confidence intervals (CIs) were employed to assess the significance of indirect effects. The results are presented in Table VI. The 95% bootstrap CIs for the indirect effects of all

four perception dimensions on customer loyalty, via customer satisfaction, excluded zero, indicating statistically significant partial mediation across all pathways. Thus, Hypothesis H6 was supported. As can be seen in Table VIII.

TABLE VIII. BOOTSTRAP MEDIATION RESULTS: INDIRECT AND DIRECT EFFECTS ON CUSTOMER LOYALTY VIA CUSTOMER SATISFACTION (H6)

Pathway	Effect Type	Effect	Boot SE	LLCI	ULCI
PU → CS →CL	Indirect	0.142	0.029	0.088	0.201
	Direct	0.279	0.051	0.180	0.379
PEOU → CS→ CL	Indirect	0.097	0.023	0.054	0.146
	Direct	0.207	0.040	0.128	0.285
ENJ → CS → CL	Indirect	0.106	0.024	0.063	0.158
	Direct	0.173	0.041	0.091	0.254
PC → CS → CL	Indirect	−0.072	0.017	−0.107	−0.042
	Direct	−0.159	0.031	−0.220	−0.098

[b] Note: Bootstrap confidence intervals not containing zero indicate significant mediation.

G. Demographic Differences in Customer Loyalty

One-way analysis of variance (ANOVA) was conducted to examine whether customer loyalty differed significantly across demographic groups. The results indicated statistically significant differences in customer loyalty by age group ($F = 3.782$, $p = 0.024$), educational level ($F = 4.563$, $p = 0.011$), and annual household income ($F = 3.136$, $p = 0.009$). No significant differences were found with respect to gender, occupation, or travel purpose. Detailed results are presented in Table IX.

TABLE IX. ONE-WAY ANOVA RESULTS: DEMOGRAPHIC DIFFERENCES IN CUSTOMER LOYALTY

Variable	Group	$\bar{x}$	SD	F	p
Age	18–29 years	3.72	0.74	3.782	0.024*
	30–49 years	3.84	0.71		
	50–60 years	3.59	0.91		
Education	High school or below	3.59	0.70	4.563	0.011*
	Bachelor's degree	3.70	0.83		
	Master's degree or above	3.89	0.69		
Annual income (RMB)	Below 30,000	3.42	0.91	3.136	0.009**
	30,000–100,000	3.73	0.83		
	100,000–300,000	3.79	0.76		
	300,000–500,000	3.76	0.70		
	Above 500,000	4.06	0.56		

[c] Note: **p < 0.01, *p < 0.05. The group with the highest mean loyalty score within each variable is shown in bold.

In summary, respondents aged 30–49 years ($\bar{x} = 3.84$), those holding a master's degree or above ($\bar{x} = 3.89$), and those with an annual household income exceeding 500,000 RMB ($\bar{x} = 4.06$) exhibited the highest levels of customer loyalty. These findings suggest that demographic characteristics play a meaningful role in segmenting hotel guests for the purpose of tailoring AI adoption strategies in hotel management to specific target groups.

H. Summary of Hypothesis Testing Results

TABLE X. SUMMARY OF HYPOTHESIS TESTING RESULTS

H	Statement	Key Statistic	Outcome
H1	The perceived usefulness of AI applications positively affects customer satisfaction.	$\beta = 0.314$, $p < 0.001$	Supported
H2	Perceived ease of use of AI applications positively affects customer satisfaction.	$\beta = 0.127$, $p = 0.008$	Supported
H3	Enjoyment of AI applications positively affects customer satisfaction.	$\beta = 0.141$, $p = 0.003$	Supported
H4	Privacy concerns regarding AI applications negatively affect customer satisfaction.	$\beta = -0.205$, $p < 0.001$	Supported
H5	Customer satisfaction positively affects customer loyalty (revisit intention and word-of-mouth).	$\beta = 0.409$, $p < 0.001$	Supported
H6	Customer satisfaction partially mediates the relationship between technological perceptions and customer loyalty.	95% Bootstrap CI excludes zero on all paths	Supported

Table X summarises the hypothesis testing results. All six hypotheses were supported at conventional significance levels. Perceived usefulness exerted the strongest positive effect on customer satisfaction ($\beta = 0.314$), followed by enjoyment ($\beta = 0.141$) and perceived ease of use ($\beta = 0.127$), while privacy concerns had a significant negative effect ($\beta = -0.205$). Customer satisfaction, in turn, positively predicted customer loyalty ($\beta = 0.409$), and bootstrap mediation analysis confirmed that it partially mediated all four perception-to-loyalty pathways, with 95% confidence intervals excluding zero across all paths.

V. DISCUSSION

A. Results Discussion

1) Perceived usefulness and customer satisfaction RQ. 1

Research Question 1: To what extent does the perceived usefulness of AI applications influence customer satisfaction among hotel guests? The regression results confirmed that PU was the strongest positive predictor of customer satisfaction ($\beta = 0.314$, $p < 0.001$), thereby supporting Hypothesis H1. This finding is consistent with Davis's [26] foundational TAM, which postulates that perceived usefulness, the degree to which a technology is believed to enhance task performance, is the primary determinant of user acceptance and satisfaction. In the present context, AI features such as the facial-recognition self-check-in terminal and the automated check-in/check-out kiosk reduce guest wait times and streamline administrative procedures. Functions that guests overwhelmingly value in an airport environment characterised by time pressure and schedule sensitivity [17]. This utilitarian primacy replicates findings reported by [12], who identified time-saving and accuracy as the most valued attributes of AI chatbots in hotel settings, and by [3], who documented that AI-customised experiences improving service efficiency exert the strongest influence on satisfaction in Chinese hotels. Furthermore, the magnitude of the PU effect in the present study ($\beta = 0.314$) is commensurate with that reported

by [5] (β = 0.29) in a comparable Chinese hotel context, suggesting that the relationship is robust across hotel tiers and city classifications. Collectively, these results reaffirm that, in functional service environments where performance outcomes are paramount, the instrumental dimension of AI perception takes precedence over hedonic and social considerations.

2) Perceived usefulness and customer satisfaction RQ. 2

Research Question 2: To what extent does the perceived ease of use of AI applications influence customer satisfaction among hotel guests? The results confirmed a significant positive effect of PEOU on customer satisfaction (β = 0.127, p = 0.008), supporting Hypothesis H2, albeit with a smaller effect size than perceived usefulness. This ordering aligns with Davis's [26] original TAM, which posited that PEOU exerts an indirect effect on acceptance primarily through its influence on PU, whereas PU retains a stronger direct path. In the hospitality context, Buhalis and Moldavska [1] demonstrated that interface complexity constitutes a significant barrier to guest adoption of voice assistant technologies, particularly among older and less technologically experienced users. The comparatively modest PEOU coefficient observed in the present study may reflect the relatively straightforward design of the facial recognition terminal, which minimises the cognitive load imposed on guests through intuitive workflow design and minimal required input. Venkatesh et al.'s [18] Unified Theory of Acceptance and Use of Technology (UTAUT) similarly found that ease of use becomes less salient as users gain familiarity with a system; given that 48.04% of respondents reported using the hotel's AI services on more than two occasions, habitual use may have attenuated the perceived difficulty of the interaction. Nonetheless, the significant positive coefficient confirms that interface simplicity remains a meaningful contributor to overall satisfaction and should not be overlooked in the design of AI service encounters.

3) Enjoyment and customer satisfaction

Research Question 3: To what extent does the enjoyment derived from AI applications influence customer satisfaction among hotel guests? Hypothesis H3 was supported, with enjoyment emerging as the second most influential positive predictor of satisfaction (β = 0.141, p = 0.003). This finding extends the hedonic dimension of the TAM framework, as elaborated by [18], into the airport hotel context, confirming that intrinsic positive affect generated by the AI interaction independently enhances satisfaction beyond utilitarian and ease-of-use considerations. The effect magnitude is noteworthy given the functional, time-sensitive nature of airport hotel stays: even in an environment that prioritises efficiency, guests derive a degree of affective value from novel and interactive AI encounters. This pattern is consistent with [3] observation that the novelty and interactivity of AI-customised hotel services engender positive emotional responses that contribute independently to satisfaction, and with [11] systematic review, which identified enjoyment as a consistent predictor of automation acceptance across tourism and hospitality contexts. The present finding is particularly significant insofar as it challenges a purely utilitarian view of AI adoption in secondary-city airport hotels; guests at Shenyang Airport Hotel, like their counterparts in more technologically mature properties, appear to respond effectively as well as cognitively to AI service encounters. This has implications for service design, suggesting that incorporating hedonic elements such as engaging interfaces and personalised recommendations can augment satisfaction even within operationally constrained AI environments.

4) Privacy concerns and customer satisfaction

Research Question 4: To what extent do privacy concerns regarding AI applications negatively influence customer satisfaction among hotel guests? The regression results confirmed that PC exerted a significant negative effect on customer satisfaction (β = −0.205, p < 0.001), thereby supporting Hypothesis H4. The magnitude of this negative coefficient is the second largest among the four predictors, underscoring the substantive role of privacy perceptions in shaping the AI service experience. This finding aligns with the theoretical arguments advanced by [19], who identified privacy unease as a consistent negative moderator of technology acceptance in hospitality settings, particularly when biometric data collection is involved. Wang and Liu [21] further documented that facial recognition-specific privacy concerns significantly depress satisfaction ratings among hotel guests, an effect amplified in transit-oriented environments such as airports, where the involuntary nature of biometric exposure is perceived as especially intrusive. The present finding that the mean privacy concern score ($\bar{x}$ = 2.98) was the only construct to fall below the scale midpoint, combined with its negative regression coefficient, indicates that privacy anxiety is a marginal concern. This trade-off dynamic, wherein guests simultaneously appreciate the convenience afforded by facial recognition whilst harbouring reservations about data collection and retention, is consistent with the ambivalence documented by [19] in their analysis of mobile payment acceptance in hotels. Addressing these concerns through transparent data governance communication is therefore not merely a compliance consideration but a direct lever for improving guest satisfaction.

5) Customer satisfaction and customer loyalty

Research Question 5: Does customer satisfaction positively influence guests' behavioural intentions, specifically revisit intention and word-of-mouth recommendation? The simple regression analysis confirmed that customer satisfaction exerted a significant positive effect on customer loyalty (β = 0.409, t = 9.501, p < 0.001), supporting Hypothesis H5 in both its revisit intention and word-of-mouth recommendation dimensions. This result is consistent with the foundational satisfaction-loyalty framework advanced by [20] and the service-quality consequences model proposed by [23], both of which posit that positive evaluative judgments of service encounters are the primary antecedent of behavioural loyalty intentions. Liu et al. [3] and Zhang et al. [5] similarly reported significant satisfaction-to-loyalty paths in AI-enabled Chinese hotel studies, reinforcing the cross-contextual stability of this relationship.

However, the model's explained variance for loyalty (R^2 = 0.167) is modest, indicating that customer satisfaction, while a necessary condition, is not sufficient to guarantee loyalty in the airport hotel context. This finding is interpretable in light of Liu and Wang's [8] structural characterisation of airport hotels as high-turnover, short-stay properties in which guests'

accommodation choices are frequently determined by flight schedules, corporate rates, and proximity to the terminal rather than by service satisfaction alone. As Zeithaml et al. [23] noted, the behavioural consequences of satisfaction vary markedly across service categories; in commodity-like, transaction-oriented settings such as airport hotels, the satisfaction-loyalty linkage is predictably weaker than in experiential or relationship-oriented segments. The present R^2 value nonetheless compares favourably with that reported by Soliman [17] ($R^2 = 0.14$) in a developing-country hotel context, suggesting that the relationship strength is consistent across settings with partial AI infrastructure. The modest explained variance (16.7%) also signals that important predictors of loyalty remain outside the present model. Three categories of omitted variables warrant attention in future research. First, perceived value guests' overall assessment of the cost-benefit ratio of the hotel stay has been identified as a direct antecedent of loyalty independent of satisfaction in multiple hospitality studies [14]. Second, trust in the hotel brand and in the AI systems themselves may constitute a partially distinct loyalty mechanism, particularly in contexts involving biometric data collection where relational confidence is a precondition for continued patronage [19, 21]. Third, structural switching costs, including corporate rate agreements, proximity to the airport, and frequency-stay benefits, are especially salient in the airport hotel segment, where accommodation choices are often constrained by logistical factors rather than purely evaluative ones [8]. Incorporating these constructs into future models would decompose the residual loyalty variance and produce a more complete and practically actionable framework.

6) Mediating role of customer satisfaction

Research Question 6: Does customer satisfaction mediate the relationships between guests' technological perceptions of AI applications and customer loyalty? Bootstrap mediation analysis (5,000 resamples; Preacher and Hayes [31]) confirmed that customer satisfaction partially mediates all four perception-to-loyalty pathways, with bias-corrected 95% confidence intervals excluding zero on all indirect effect estimates. Hypothesis H6 was therefore supported. The partial mediation pattern wherein direct effects from all four perception dimensions on loyalty remain significant alongside the significant indirect effects through satisfaction has important theoretical implications.

First, the findings indicate that technological perceptions shape loyalty through two complementary mechanisms: an evaluative route, whereby favourable perceptions enhance satisfaction and satisfaction subsequently drives loyalty intentions, and a direct attitudinal route, whereby favourable AI perceptions independently generate loyalty predispositions that bypass the satisfaction evaluation. This dual-pathway structure is consistent with the theoretical extensions to TAM proposed by [18], who argued that technology beliefs influence behavioural intentions both through affective and evaluative intermediaries and through direct cognitive appraisals. Second, the finding that all four perception dimensions, including the negative pathway through privacy concerns, exhibit partial mediation through satisfaction suggests that satisfaction is a robust but not exhaustive transmission mechanism. Zhang et al. [5] similarly documented partial mediation by satisfaction in a Chinese hotel AI study, and Liu et al. [3] observed that the AI-customisation-to-loyalty path retained direct components even after controlling for satisfaction, corroborating the present pattern. Third, the indirect effect of perceived usefulness through satisfaction (0.142) is the largest among the four indirect pathways, consistent with its dominant direct effect, confirming that utility-driven satisfaction has the most substantial indirect impact on loyalty. By contrast, the negative indirect effect of privacy concerns (−0.072) is the smallest in absolute magnitude, suggesting that while privacy anxiety depresses satisfaction, its downstream impact on loyalty is partially buffered by the direct positive effects of usefulness and enjoyment. These mediation results collectively affirm that improving guest satisfaction through AI service optimisation constitutes a strategically effective, if not singular, pathway to enhancing customer loyalty in the airport hotel context.

7) Demographic differences in customer loyalty

One-way ANOVA revealed significant inter-group differences in customer loyalty by age ($F = 3.782$, $p = 0.024$), educational level ($F = 4.563$, $p = 0.011$), and annual household income ($F = 3.136$, $p = 0.009$), while gender, occupation, and travel purpose yielded no significant differences. Guests aged 30–49 years exhibited the highest mean loyalty score ($\bar{x} = 3.84$), followed by those holding postgraduate qualifications ($\bar{x} = 3.89$) and those with annual household incomes exceeding 500,000 RMB ($\bar{x} = 4.06$). These patterns are consistent with [11] finding that actual digital experience and cognitive engagement with technology, rather than chronological age per se, are the more proximate determinants of AI acceptance; the 30–49 cohort's higher loyalty likely reflects greater professional familiarity with AI interfaces rather than generational disposition alone. The absence of elevated loyalty among the youngest respondents (18–29 years) challenges the commonly invoked assumption that digital natives automatically embrace all forms of AI-mediated service, a supposition that [9] cautioned against in their review of age as a technology acceptance moderator. The education and income gradients align with [17] finding that higher socioeconomic status is associated with greater receptiveness to AI in hotel settings, potentially because highly educated and higher-income guests are more accustomed to interacting with digital systems and interpret AI features as quality signals rather than sources of anxiety. For managerial practice, these demographic differentiations imply that a uniform AI strategy is unlikely to optimise loyalty across all guest segments.

VI. Conclusion

Drawing on an integrated TAM and Customer Experience Theory framework, this study examined how guests' technological perceptions of AI applications influence satisfaction and loyalty at Shenyang Airport Hotel. Data from 452 guests confirmed that perceived usefulness, perceived ease of use, and enjoyment positively affect customer satisfaction, while privacy concerns exert a significant negative effect, supporting H1–H4. Customer satisfaction in turn positively predicts customer loyalty (H5) and partially mediates all four perception-to-loyalty pathways (H6). Demographic analysis further revealed significant differences in loyalty across age,

educational level, and income groups, but not across gender, occupation, or travel purpose.

These findings collectively demonstrate that, even within a technologically constrained secondary-city airport hotel, guests' multidimensional AI perceptions systematically shape satisfaction and loyalty outcomes. For practitioners, the results highlight the importance of maximising perceived utility, simplifying service interfaces, enriching hedonic engagement, and communicating data governance policies transparently. For researchers, this study extends TAM into an underexamined hospitality context and provides rigorous mediation evidence clarifying the perception satisfaction loyalty mechanism. Ultimately, effective AI adoption in hotel management requires attending simultaneously to utilitarian, affective, and privacy-related perceptions to convert technological investment into sustainable guest retention.

VII. Research Implications

1) Theoretical implications: This study makes three principal theoretical contributions. First, it extends the TAM [26] beyond its original workplace and basic utilitarian boundaries by integrating both an affective facilitator (enjoyment) and a risk-based barrier (privacy concerns) into a single model applied to AI-mediated guest services in a secondary-city airport hotel. Prior research in the hospitality domain had typically addressed these constructs in isolation: [12] examined PU and PEOU without incorporating hedonic or privacy dimensions, while [16] and [23] focused exclusively on privacy concerns. The present simultaneous treatment of all four perception dimensions demonstrates that both affective and security-related factors play statistically significant and conceptually distinct roles in satisfaction formation, confirming that a purely utilitarian view of AI acceptance is insufficient in service contexts where emotional responses and risk perceptions are salient, as argued by [18] and [11].

Second, the bootstrap mediation results provide rigorous empirical evidence of partial mediation through satisfaction, clarifying the perception-to-loyalty transmission mechanism that has remained theoretically postulated but empirically underexamined in prior work [3, 20, 22]. By demonstrating that direct perception-to-loyalty paths coexist with satisfaction-mediated indirect paths across all four constructs, this study contributes a more nuanced causal architecture to the existing satisfaction-loyalty literature [20, 23]. Future research incorporating additional mediators — such as perceived value or trust, as suggested by [14] could further decompose the residual direct effects.

Third, by focusing on an airport hotel in a secondary Chinese city, a context characterised by partial AI infrastructure, low CRM maturity, and an annual repurchase rate of 18%, this study addresses a significant gap in a literature dominated by first-tier urban luxury properties [13, 16]. The finding that TAM-based perceptions are systematically related to satisfaction and loyalty even in technologically constrained settings implies that the framework is transportable across varying levels of AI infrastructure maturity. This aligns with [17] demonstration that TAM relationships hold in developing-country hotel contexts, and with [10] multi-case study findings from Thailand, collectively suggesting that the model's explanatory power is not contingent upon sophisticated technological ecosystems.

2) Practical implications: The findings carry several actionable implications for hotel managers, particularly those operating three-star airport properties with limited technological resources. First, as perceived usefulness emerged as the dominant driver of satisfaction, investment in AI features that deliver clear, measurable time savings should be prioritised. Automated check-in/check-out and facial recognition terminals must be robust, fast, and error-free, as any friction in these high-stakes interactions directly erodes perceived utility [12, 18, 26]. Second, interface simplicity should be treated as a design principle rather than a secondary consideration. Buhalis and Moldavska [9] demonstrated that clear instructions, logical workflow sequences, and multilingual support are low-cost interventions that substantially improve perceived ease of use for older and less technologically experienced guests; such enhancements are within reach even for properties operating without dedicated IT personnel, as is the case at Shenyang Airport Hotel.

Third, the significant enjoyment effect indicates that augmenting AI interactions with hedonic elements — such as gamified loyalty point accrual, personalised welcome messages, or voice-interactive features- can strengthen satisfaction without requiring substantial capital expenditure. Liu et al. [3] and Tussyadiah [11] both emphasised that guests respond to novelty and interactivity, suggesting that incremental enhancements to engagement design can yield disproportionate satisfaction gains. Fourth, the negative effect of privacy concerns represents the most urgent operational priority. Hotels deploying facial recognition should provide concise, comprehensible explanations of what biometric data are collected, the duration of retention, and the parties with access, following the transparency communication model advocated by [19] and [21]. Offering a privacy-preserving opt-out alternative for guests who prefer conventional check-in may reduce anxiety without disrupting service continuity for the majority.

Finally, the ANOVA-derived demographic differences in loyalty suggest that a one-size-fits-all AI deployment strategy is suboptimal. In particular, the higher loyalty reported by postgraduate and high-income guests indicates receptiveness to personalised, data-driven service enhancements such as AI-driven loyalty programmes and recommendation engines — features that Makivic et al. [13] and Chen [16] identified as particularly effective loyalty drivers in technologically sophisticated hotel segments. Conversely, the lower loyalty among the youngest and lowest-income cohorts signals a need for simplified, guided AI interactions supplemented by strong privacy assurances, in line with the segmentation principles advocated by [14].

VIII. Limitations and Future Work

Several limitations of this study merit acknowledgment. First, the cross-sectional survey design permits the examination of associations among constructs at a single point in time but precludes the establishment of temporal causality. Longitudinal

designs tracking the same guests across multiple stays would more robustly establish the causal direction and stability of the perception-satisfaction-loyalty relationships documented here.

Second, data were collected from a single three-star airport hotel in Shenyang, a deliberate design choice that enables contextual depth but constrains generalisability. The findings should not be generalised uncritically to other hotel categories (four- or five-star properties, boutique hotels, or resort segments), to first-tier Chinese cities with more mature AI infrastructure, or to international hospitality markets with different regulatory and cultural orientations toward biometric data collection. Future research should replicate the model across multiple airport hotels in both secondary and primary Chinese cities, and potentially in comparable international contexts such as the Thai airport hotel setting examined by [10], to assess whether the TAM-based relationships hold under varying levels of AI infrastructure maturity and cultural orientation.

Third, the study relied exclusively on self-reported Likert-scale measures, which may be subject to common method variance and social desirability bias despite the procedural safeguards implemented [32]. No objective behavioural measures—such as actual repeat bookings, transaction records, or AI usage logs—were included to validate the reported loyalty intentions, which limits the correspondence between stated and enacted behaviour. Integrating objective behavioural data, including verified repurchase records from the hotel's reservation system, actual check-in duration logs from the facial recognition terminal, and transaction-level spending data, with perceptual survey data in future work would substantially strengthen causal inference, mitigate common method bias, and provide a more complete picture of the satisfaction-loyalty conversion process. Furthermore, the study did not control for potentially confounding factors such as room quality, service quality beyond AI interactions, room pricing, or locational convenience (e.g., proximity to the terminal and shuttle availability), all of which may independently influence satisfaction and loyalty and whose omission may have inflated the estimated effects of AI perceptions. Future studies should include these covariates, or employ matched-pair designs that hold non-AI service quality constant, to isolate the incremental contribution of AI perceptions more rigorously.

Fourth, the conceptual model did not include theoretically plausible moderating variables. Venkatesh et al. [18] demonstrated that prior experience, technology anxiety, and cultural orientation moderate TAM relationships; in the present context, guests from different national backgrounds may hold systematically divergent privacy expectations regarding biometric data collection, a moderating pathway unexplored in this study. Morosan and DeFranco [19] similarly called for research examining how cultural and individual difference variables shape the privacy-acceptance dynamic in hospitality AI deployments. Future studies incorporating these moderators would produce a more complete and generalisable model of AI perception formation. In particular, technology readiness [18], prior AI experience, frequency of hotel stays, and individual-level trust dispositions each represent theoretically motivated moderators that may alter the strength or direction of the perception-satisfaction and satisfaction-loyalty relationships identified in this study. Incorporating such variables would clarify boundary conditions and enhance the practical precision of the model.

Finally, this study did not examine the cost-effectiveness of different AI implementations, an omission that limits its managerial utility in the context of resource-constrained properties such as Shenyang Airport Hotel. Integrating economic evaluation frameworks such as return-on-investment analysis for each AI feature with the perceptual satisfaction and loyalty models developed here would provide hotel managers with the complete evidence base needed to prioritise AI investments strategically. This represents a productive and practically significant direction for future inquiry.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by Suratthani Rajabhat University. Ethical approval for this study was granted by the Research Ethics Committee of Suratthani Rajabhat University. The authors also extend their sincere appreciation to all participants, subject-matter experts, and individuals who generously contributed their time and insights to this research.

REFERENCES

[1] D. Buhalis and R. Leung, "Smart tourism destinations: Ecosystems for tourism destination competitiveness," Int. J. Tourism Res., vol. 20, no. 4, pp. 411–421, 2018.

[2] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins, "Brave new world: Service robots in the frontline," J. Service Manage., vol. 29, no. 5, pp. 907–931, 2018.

[3] H. Liu, Y. Li, Y. Hu, and C. Wu, "The influence of AI-customized experiences on customer satisfaction and loyalty in the hotel industry," Int. J. Hospitality Manage., vol. 92, p. 102746, 2021.

[4] S. Ivanov and C. Webster, "Adoption of robots and service automation by tourism and hospitality companies," Revista Turismo & Desenvolvimento, vol. 31, pp. 9–29, 2019.

[5] L. Zhang, X. Wu, and M. Chen, "The impact of AI-driven personalization on hotel performance: Evidence from China," J. Service Res. China, vol. 10, no. 1, pp. 67–82, 2022.

[6] Shenyang Airport Hotel Management, "Operational statistics and hotel background briefing," personal communication, Mar. 2026.

[7] Authors' preliminary field survey, "Preliminary field survey of Shenyang Airport Hotel: Operational performance and AI infrastructure assessment," unpublished raw data, Suratthani Rajabhat University, Apr. 2026.

[8] Y. Liu and L. Wang, "Integrated operations of airport hotels: Synchronizing with transportation hubs," J. Tourism Oper. Manage., vol. 18, no. 2, pp. 134–148, 2023.

[9] D. Buhalis and I. Moldavska, "Voice assistants in hospitality: Using artificial intelligence for customer service," J. Hospitality Tourism Technol., vol. 13, no. 3, pp. 386–403, 2022.

[10] N. Chotisarn, "Smart hotel technology adoption in Thailand: A multi-case study," J. Asian Tourism Res., vol. 12, no. 1, pp. 44–61, 2025.

[11] I. P. Tussyadiah, "A review of research into automation in tourism," Ann. Tourism Res., vol. 81, p. 102883, 2020.

[12] J. Shin and E. Jeong, "Perceived usefulness and ease of use of AI chatbots in hotels," J. Hospitality Marketing Manage., vol. 29, no. 4, pp. 435–452, 2020.

[13] I. Makivic, B. Makivic, D. Nikolic, and V. Jovanovic, "AI-powered guest experience in luxury hotels: A systematic review," J. Quality Assurance Hospitality Tourism, vol. 25, no. 3, pp. 298–319, 2024.

[14] J. Bulchand-Gidumal, E. William Secin, P. O'Connor, and D. Buhalis, "Artificial intelligence's impact on hospitality and tourism marketing," Current Issues Tourism, vol. 27, no. 14, pp. 2345–2362, 2024.

[15] Marriott International, 2024 Sustainability and Social Impact Report. Bethesda, MD: Marriott International Inc., 2024.

[16] X. Chen, "AI-driven service personalization and guest loyalty in Chinese hotel chains," Int. J. Hospitality Manage., vol. 115, p. 103628, 2025.

[17] M. Soliman, "Adoption of AI-driven personalization in hotel services: A developing-country perspective," Tourism Manage. Perspect., vol. 49, p. 101198, 2024.

[18] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quart., vol. 27, no. 3, pp. 425–478, 2003.

[19] C. Morosan and A. DeFranco, "Modeling guests' intentions to use mobile apps in hotels," Int. J. Contemporary Hospitality Manage., vol. 28, no. 9, pp. 1968–1991, 2016.

[20] R. L. Oliver, Satisfaction: A Behavioral Perspective on the Consumer. New York: McGraw-Hill, 1997.

[21] H. Wang and Q. Liu, "Privacy concerns and guest satisfaction with AI in hotels," J. Hospitality Law Ethics, vol. 15, no. 2, pp. 145–162, 2023.

[22] P. C. Verhoef, K. N. Lemon, A. Parasuraman, A. L. Roggeveen, M. Tsiros, and L. A. Schlesinger, "Customer experience creation: Determinants, dynamics and management strategies," J. Retailing, vol. 85, no. 1, pp. 31–41, 2009.

[23] V. A. Zeithaml, L. L. Berry, and A. Parasuraman, "The behavioral consequences of service quality," J. Marketing, vol. 60, no. 2, pp. 31–46, 1996.

[24] J. W. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 4th ed. Thousand Oaks, CA: Sage Publications, 2014.

[25] W. G. Cochran, Sampling Techniques, 3rd ed. New York: John Wiley & Sons, 1977.

[26] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quart., vol. 13, no. 3, pp. 319–340, 1989.

[27] R. W. Brislin, "Back-translation for cross-cultural research," J. Cross-Cultural Psychol., vol. 1, no. 3, pp. 185–216, 1970.

[28] R. J. Rovinelli and R. K. Hambleton, "On the use of content specialists in the assessment of criterion-referenced test item validity," Dutch J. Educ. Res., vol. 2, pp. 49–60, 1977.

[29] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis, 8th ed. Mason, OH: Cengage Learning, 2019.

[30] H. F. Kaiser, "An index of factorial simplicity," Psychometrika, vol. 39, pp. 31–36, 1974.

[31] K. J. Preacher and A. F. Hayes, "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models," Behav. Res. Methods, vol. 40, no. 3, pp. 879–891, 2008.

[32] P. M. Podsakoff, S. B. MacKenzie, J. Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," J. Appl. Psychol., vol. 88, no. 5, pp. 879–903, 2003.

Ontology-Based Business Process Modeling: A Review

Low Kok Thai[1], Furkh Zeshan[2*], Nazri Kama[3*], Riza Sulaiman[4], Mohammad Nazir Ahmad[5], Uwais Qidwai[6]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia (UKM), Malaysia[1, 2, 4, 5]
Department of Computer Science, COMSATS University Islamabad (CUI), Lahore Campus[2]
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia (UTM), Malaysia[3]
Department of Computer Science & Engineering, Qatar University[6]

***Abstract*—Business Process Modeling (BPM) has been receiving attention in recent years. Organizations operating in distributed, data, and knowledge-intensive environments need precise machine-interpretable process descriptions. Traditional business process modeling notations such as BPMN, UML Activity Diagrams, and EPCs are highly effective for visualizing workflows and supporting communication among stakeholders but do not address problems such as semantic ambiguity, inconsistent terminology, limited reuse, and poor interoperability across organizational boundaries. Since ontology-based models can enable seamless process integration, coordination, and collaboration among autonomous systems. Therefore, ontology-based descriptions are well-suited for complex enterprise systems and supply chains. Ontology-based business process modeling techniques provide a robust and theoretically grounded solution by introducing explicit semantics, formal reasoning capabilities, and shared conceptual frameworks into business process models. This semantic enrichment significantly enhances the expressive power of business process descriptions while preserving compatibility with existing modeling standards. However, ontology-based business process modeling also has its own challenges. The development and maintenance of high-quality ontologies require significant effort and domain expertise. Moreover, reasoning over large-scale ontologies may introduce computational overhead, particularly in real-time environments. Tool support is another practical concern, as seamless integration between business process modeling tools and ontology management platforms is still limited in many industrial settings. The primary objective of this study is to conduct a systematic literature review of ontology-based business process modeling approaches to provide research recommendations based on their strengths and limitations. The results indicate that there are several research gaps that should be addressed to ensure smooth process integration across organizational boundaries. Additionally, empirical validation of ontology-based BPM frameworks in real-time environments is limited. The proposed framework not capable enough to update ontologies with the evolution of the business and related processes. In this regard, an Input Process Output (IPO) BPM Framework to integrate ontologies into BPM as a three-stage transformation mechanism is proposed.**

***Keywords*—Business Process Modeling (BPM); Ontology-based Business Process Modeling (OBPM); domain modeling; knowledge management; knowledge reasoning; knowledge engineering; knowledge modeling**

*Corresponding author.

I. Introduction

Business Process Modeling (BPM) deals with discovering knowledge from information technology (IT) and management science. Organizations analyze, model, and optimize the acquired knowledge to increase the process's efficiency, reduce operational costs, and improve customer satisfaction [1 - 3].

In BPM, the processes are described as structured or semi-structured, but related activities are executed in series or parallel to ensure the proper operations are performed in an organization. Whereas the process model usually describes the temporal activities performed specifically for the creation and use of data [4]. Business processes (BP) in most cases represent complex systems where several entities are involved in their development. For example, in the case of billing a car, nine entities $\langle a, c, e, f, g, i, j, k, l \rangle$ may be involved, starting from booking a car (activity a), paying extra insurance (activity c), booking confirmation (activity e), the check-in process (activity f), car selection (activity i), checking license (activity j), charging credit card (activity k), and finally car releasing (activity l). However, the main aspects of business processes are the activity sequencing, where activities are performed in a specific order to achieve a desired result, and a workflow, in which activities are often connected in such a way that the output of one activity becomes the input of the next [4]. However, generating process models from unstructured text and implementing them in an organization is a challenging task. Although the advent and the application of Natural Language Processing (NLP) have made this process easier, natural language still presents a significant challenge due to the inherent ambiguity, semantic nuance, sentence composition, word relevance, and contextual connections between sentences [5]. Moreover, the business process modelers may misunderstand domain concepts or relationships among these concepts due to their lack of precise domain knowledge. In such cases, semantic ambiguity affects the efficiency and quality of business process modeling. To address this problem, ontology-based business process modeling techniques are the solution, which can ease semantic ambiguity by providing a means to capture rich, semantic information on complex business processes through domain-specific ontologies.

Business process modeling is a knowledge-intensive activity where modelers meet knowledge workers, business owners, and IT professionals to gain an accurate understanding of domain knowledge and process logic [6 - 9]. However, if the domain

becomes larger and more complex, then an accurate understanding of the domain can be difficult for a modeler. Particularly if the modeler does not belong to the domain, such situations may lead to semantic ambiguity (precise mapping between domain concepts will be difficult). Semantic ambiguity, as a result, may lead to high cognitive load on the modeler, resulting in over- or underspecification, thus reducing the quality of a process model [10 - 13].

A process model (conceptual model) defines how a business will conduct its operations by specifying relevant activities, events, states, and control flow logic [14, 15]. These models can be represented with Petri Nets [16, 17], Business Process Model and Notation as shown in Fig. 1 (BPMN) [18], and Event-driven Process Chain (EPC) [19] shown in Fig. 2. Although these formal methodologies to model business processes offer different perspectives on modeling, they still do not address semantic ambiguity problems directly [20].

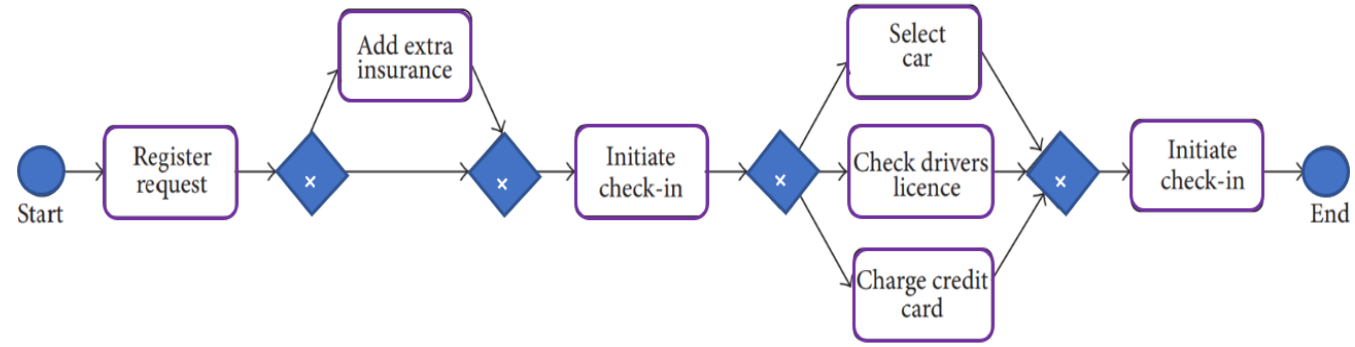


Fig. 1. Model in business process modeling notation.

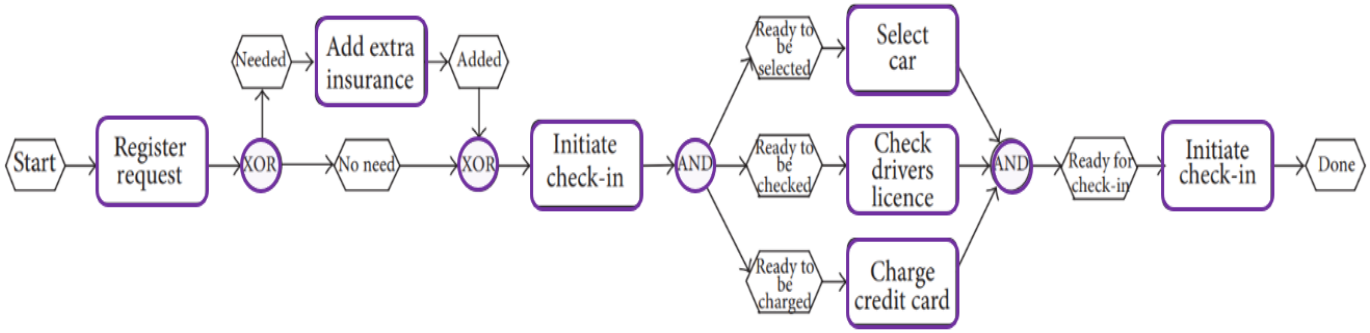


Fig. 2. Event-driven process chain model.

However, ontology-based business process modeling approaches can resolve semantic ambiguity in business process modeling through domain ontologies. Such approaches are based on cognitive load theory [21] and intended to help reduce semantic ambiguity by avoiding problems such as construct overload and construct redundancy. Construct overload deals with modeling for two or more ontologies, while construct redundancy deals with modeling constructs for a single ontology. Moreover, classes in a domain process ontology formally guide the business process modeling.

The primary objective of this study is to conduct a review of ontology-based business process modeling approaches for three main reasons:

- to establish context and understanding of the current state of the art knowledge;
- to explore the unexplored areas and to identify the research gaps that should be addressed; and
- to identify research opportunities for future research.

In this regard, a Systematic Literature Review (SLR) is performed because, according to Barbara Kitchenham [22], SLR is a systematic method for identifying, evaluating, and synthesizing research evidence that follows a structured protocol to ensure comprehensiveness and minimize bias.

The rest of the study is organized as follows: Section II presents the background of the study. Section III presents the review of ontology-based business process modeling. In Section IV, research challenges and opportunities are discussed, and finally, in Section V, the conclusion and prospective future research areas are discussed.

II. Background

A. What is BPM?

Business process management (BPM) is an approach to improve the activities to get work done, satisfy customer needs, and generate business value. It includes concepts, methods, and techniques to support the design, administration, configuration, enactment, and analysis of business processes [2, 23, 24]. To assess the effectiveness of implemented BPM in any organization, different factors can be used; a few among them are the process owner's awareness, engagement of top management, close communication, linking process goals with personal as well as strategic goals, the degree of innovation,

corrective and preventive measures, collaborative environment, process documentation, etc.

Fig. 3 depicts the core activities of business process management. Business process modelling is the core phase of process design. Once an initial design of a business process is developed, it needs to be validated.

Fig. 4 shows the BPM reference architecture [26] where the designer uses the design tools to create models describing the processes and the structure of the organization. The manager uses management tools to monitor the flow of work and act if necessary. The worker interacts with the enactment service.

Fig. 3. BPM life cycle core activities [25].

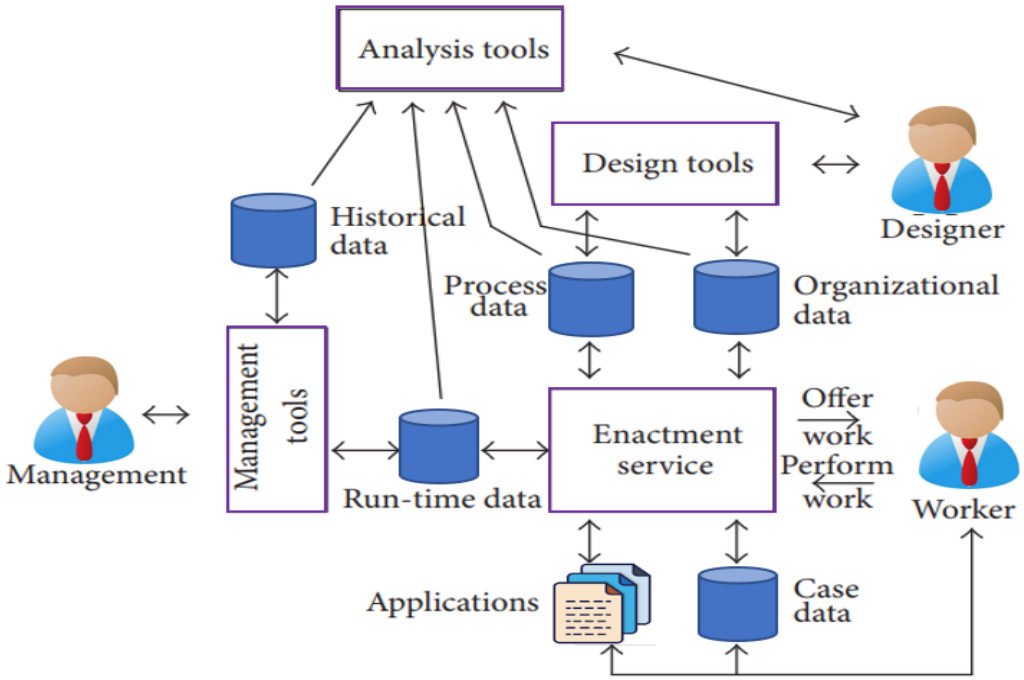


Fig. 4. BPM reference architecture.

B. What is an Ontology?

Ontology is the philosophical study of being, existence, and reality, which investigates what kinds of things exist and how they are categorized. In the context of computer science and information science, it refers to a formal, explicit specification of a conceptualization, essentially a structured representation of a domain's concepts and their relationships. A generic ontology overview is presented in Fig. 5.

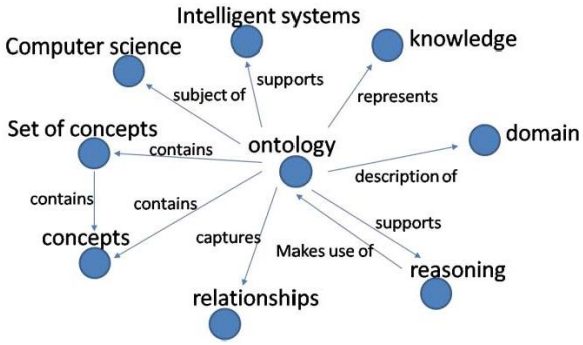


Fig. 5. Generic overview of an ontology.

Key issues in ontology-based business process modeling include the complexity of building and maintaining ontologies, lack of clear semantics in traditional modeling tools, the potential for mismatches between modeling approaches and technological trends, and the need for improved methods to capture complex business rules. Additionally, there can be difficulties in achieving interoperability and handling dynamic, complex business environments.

C. Why BPM Leverages Ontology?

Business process (BP) models are conceptual models; when implemented, they create value for the underlying business. This is why companies are focusing on defining and refining their business processes. However, BP modeling is a difficult task that needs trained experts. Although business process models are very similar to conceptual models, they are much more complex [12]. Because these models are used to draw the dynamic behavior of the business, that's why it requires a good understanding of domain knowledge. Whereas domain knowledge in the Information Systems (IS) discipline refers to the knowledge provided by both the domain experts and the application repositories [27]. Moreover, the next generation of

BPM requires machine-understandable semantics to automate the process composition, mediation, and execution. But the lack of explicit semantics hinders this task [28]. To address this challenge, different semantic modeling techniques like ontologies, semantic annotation, knowledge graphs, and rule-based systems can be used. Semantic annotation is good to ensure the quality of the model, but manual annotation is prone to errors and time-consuming [57]. Whereas in the case of Knowledge Graphs, knowledge extraction and integration from heterogeneous sources in complex scenarios is very difficult [55] while in Rule-Based Systems, combining rule logic with graph control flow may lead to modeling inconsistencies in business logic [56]. In contrast, ontologies can play their role between business process modeling and technical implementations by defining concepts, their attributes, and relations formally [20]. Hence, enabling the process models to be understood and manipulated by software systems, resulting in a higher degree of automation and execution.

D. Issues and Challenges While Integrating Ontologies Into BPM

The basic aim of integrating ontologies into BPM is to enhance semantic interoperability, formalize process models, and enable automated reasoning, but it faces several technical, conceptual, and practical hurdles. A few key challenges are discussed below and presented graphically in Fig. 6. The first challenge is that creating and maintaining a comprehensive ontology that covers all business domains, rules, and processes is very difficult. Moreover, with the evolution of the business and related processes, updating the ontology is another challenging task [29]. Different businesses may use different terms for the same concept, leading to the semantic mismatch problem. To resolve such issues, aligning heterogeneous ontologies is another major technical problem [30].

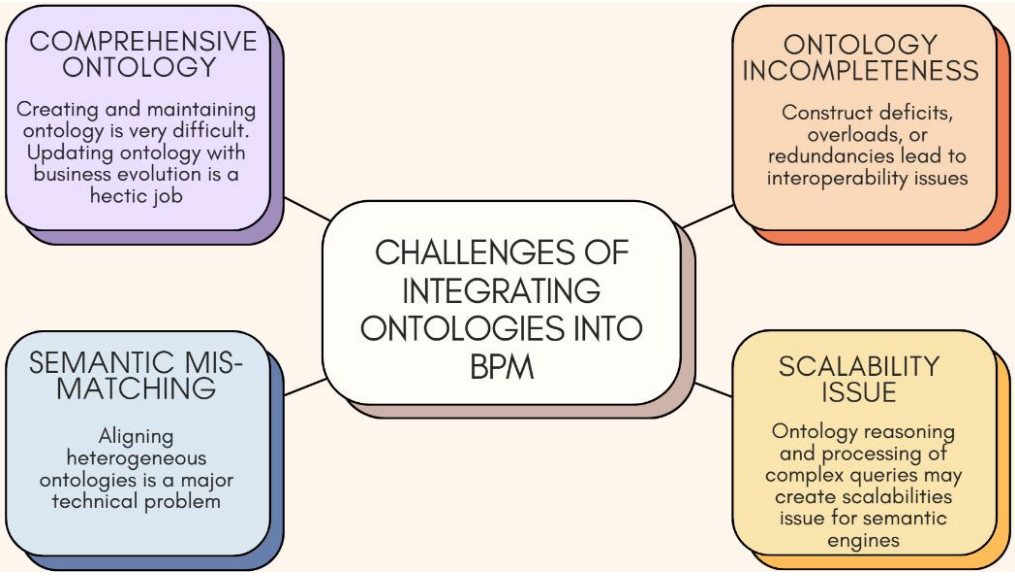


Fig. 6. Ontology integration challenges into BPM.

Ontological incompleteness (missing a crucial type of object or relationship) may lead to construct deficits, overloads, or redundancies, which undermine clear semantic mappings and interoperability [31]. Such issues arise because BPMN has no way to formally model the rules, resulting in writing them in text on the diagram or in a separate document. The ontology of business rules is not complete within BPMN [32]. Ontology reasoning, consistency checking, and processing of complex queries in the presence of a high volume of BPM scenarios may create scalability issues for the semantic engines [33].

There is a gap between conceptual ontological models and their implementation through BPM tools that support both process modeling and ontology development [31]. Developing and adopting standardized, integrated methods and tools should be a focus.

E. Ontology-Based BPM Framework

With the ever-increasing complexity of evolving businesses, there is a need for organizations to continuously refine, redefine, and formalize their business processes to create maximum value. However, business process modeling for large businesses is a complex and time-consuming activity that requires deep domain expertise. Moreover, the absence of explicit machine-understandable semantics in traditional BPM tools poses limitations in automation, interoperability, and process intelligence. To address these limitations, ontologies can play their role by offering means to formalize the specification of domain concepts, relationships, and rules, resulting in enhanced semantic interoperability and automated reasoning. Hence, researchers are exploring the integration of ontologies into BPM [30, 31, 33].

However, the integration of domain ontologies into BPM has numerous challenges, such as semantic heterogeneity (different organizations use different terms for the same concept), ontology evolution, maintaining comprehensive ontologies, reasoning scalability, and ontological incompleteness (missing relationships or business rules), etc. These limitations highlight the need for a structured framework that systematically links the semantic and procedural dimensions of BPM. In this regard, an Input Process Output (IPO) BPM Framework (Fig. 7) to

integrate ontologies into BPM as a three-stage transformation mechanism is needed. At the input stage, the framework will capture domain knowledge, ontological definitions, and business rules from both experts and organizational repositories. In the second stage (Process), semantic alignment, reasoning, and the transformation of conceptual models into machine-interpretable forms will be performed. Finally, the output stage will generate executable, semantically enriched process models that will support automation, consistency checking, and interoperability.

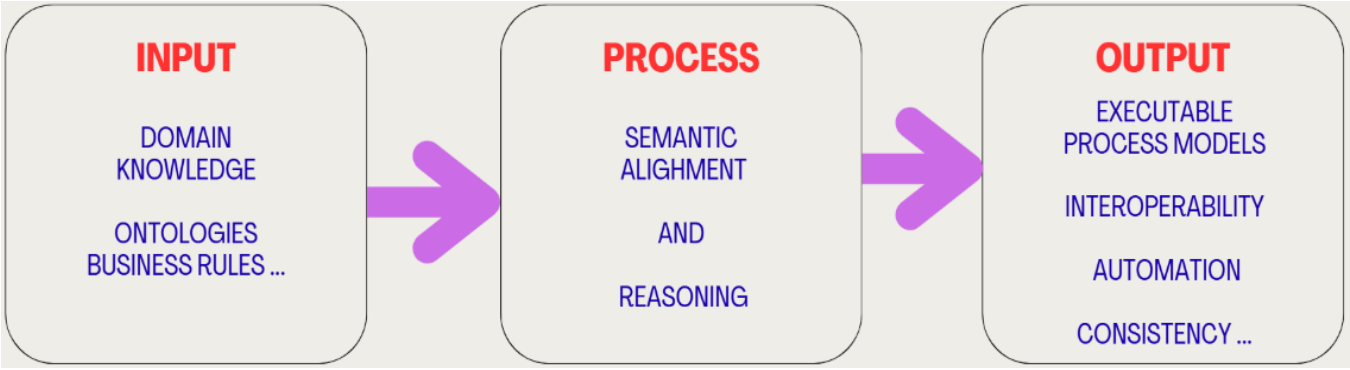


Fig. 7. Ontology-based conceptual IPO framework for BPM derived from literature.

An ontological framework where BPM integration structures around the IPO model may have the ability to ensure traceability and modularity across semantic and procedural layers. Thus, the ontology-based IPO framework for BPM not only addresses the technical and conceptual limitations of current BPM models but also contributes to establishing a unified foundation for semantic-driven business process automation.

III. Review Methodology

Ontology-based business process modeling (OBPM) uses an ontology to provide a formal, shared vocabulary for defining business processes, which allows for greater automation, reasoning, and interoperability. It integrates domain knowledge, process activities, and organizational context into a semantic model, enabling tasks like semantic gap analysis and automated metric computation that are difficult with traditional methods. While promising, the widespread adoption of OBPM is currently limited by the manual effort required for ontology creation and annotation.

A. Research Questions

The research question defines the scope of the research along with the specific issues to be investigated. Moreover, without such questions, the reader may face problems in understanding the rationale of the research. Therefore, the following are the research questions defined for this study.

- What challenges or limitations are reported in ontology-based BPM research?
- What ontological types are required to represent BPM effectively?
- How can OBPM support interoperability between heterogeneous business organizations?
- How have ontology-based BPM approaches been evaluated or validated in existing studies?

The rationale of each research question is presented in Fig. 8.

Fig. 8. Rationale of research questions.

Review protocol validation is an important step to minimize bias (in study selection, data extraction, and analysis) and to ensure transparency (allowing other researchers to replicate the methodology and to verify and validate the findings) of a review's findings. Therefore, a well-defined and validated protocol can guide the researcher throughout the review process and promote methodological rigor. In this regard, a review protocol validation questionnaire is defined in Table I.

TABLE I. REVIEW PROTOCOL VALIDATION QUESTIONNAIRE

Sr.	Components	Questions
1	Background	Is the review's context and motivation clearly established?
2	Research questions	Are research questions focused specifically on ontology-based business process modeling?
3	Search strategy	Is the completeness of searching of relevant studies ensured?
4	Study Selection	Are inclusion and exclusion criteria clearly defined?
5	Quality of studies	Are the reliability and validity of the selected studies evaluated and justified?
6	Data extraction	Is the collected information from each study relevant, consistent, and in a structured form?
7	Data synthesis	Is the extracted data integrated and analyzed properly to answer the research questions?
8	Reporting	Are findings presented clearly, ensuring others can reproduce the same if replicated?

B. Search Strategy

The search strategy is an important component of a systematic literature review methodology, which determines the scope and comprehensiveness of the review. This strategy ensures that all relevant studies are identified without being biased. Moreover, it also serves as a tool for evaluating the quality of a search. It allows replication upon updating a review. In this regard, a search strategy for this survey is designed in Fig. 9.

A search string is a combination of keywords and operators (like AND & OR) to form and perform complex, targeted queries and filter relevant results crucial for research. These strings are used for searching the databases to find specific information. A well-defined string makes the search more precise than single words. Below are the search strings used in this research.

- (("Ontology" OR "Ontologies" OR "Ontological") AND ("Business Process" OR "Business Process Model"))
- ("Business Process" OR "Business Process Model" OR "Business Process Management" OR "BPM")
- (("Ontology-based" OR "Ontology-driven" OR "Ontology based" OR "Ontology driven") AND ("ontology supported" OR "Semantic"))
- ("Business Process Model" OR "BPMN" OR "OBPM")

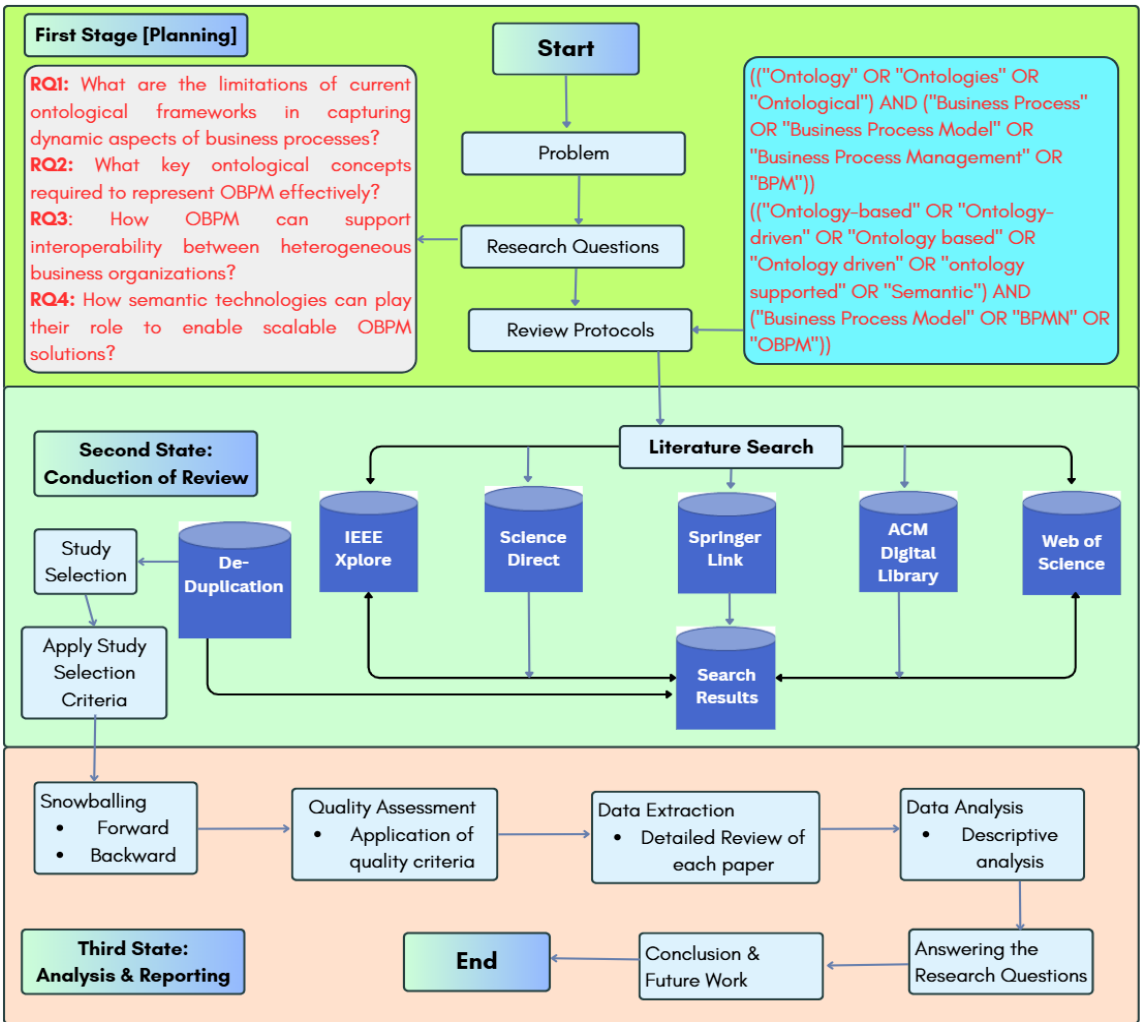


Fig. 9. Literature research methodology.

Databases provide a collection of published research for review on any topic. To search relevant papers, a researcher can define and refine the search filters to ensure only peer-reviewed articles are available. In order to ensure that no relevant study is missing, a researcher must search multiple databases. In this regard, below is the list of databases that are searched for literature for this review.

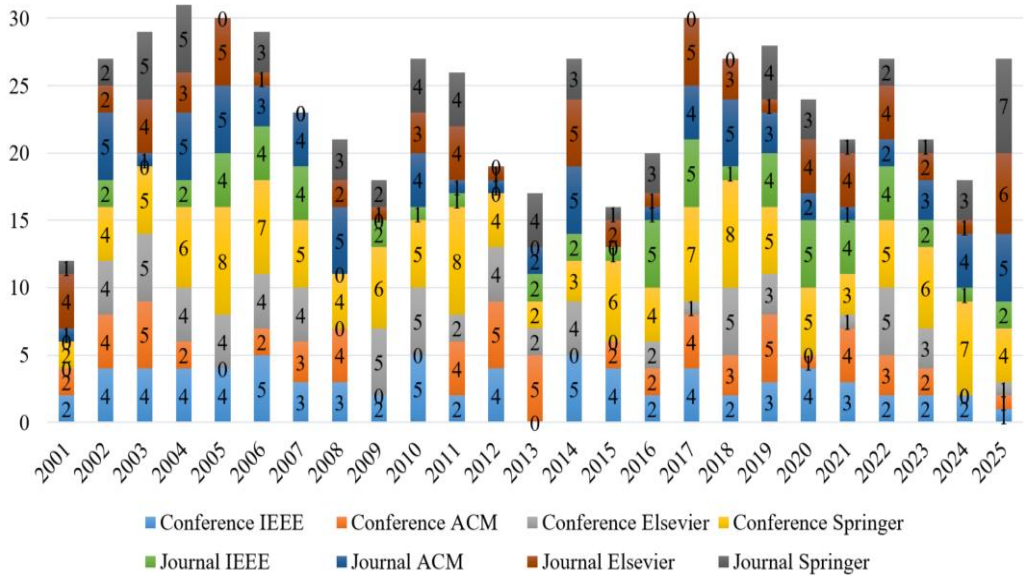
- ACM Digital Library (https://dl.acm.org/)
- IEEE Explore Digital Library (https://ieeexplore.ieee.org/)
- Springer Link Online Library (https://link.springer.com/)
- Elsevier ScienceDirect (https://www.sciencedirect.com/)
- Web of Science (https://www.webofscience.com)

C. Study Selection Criteria

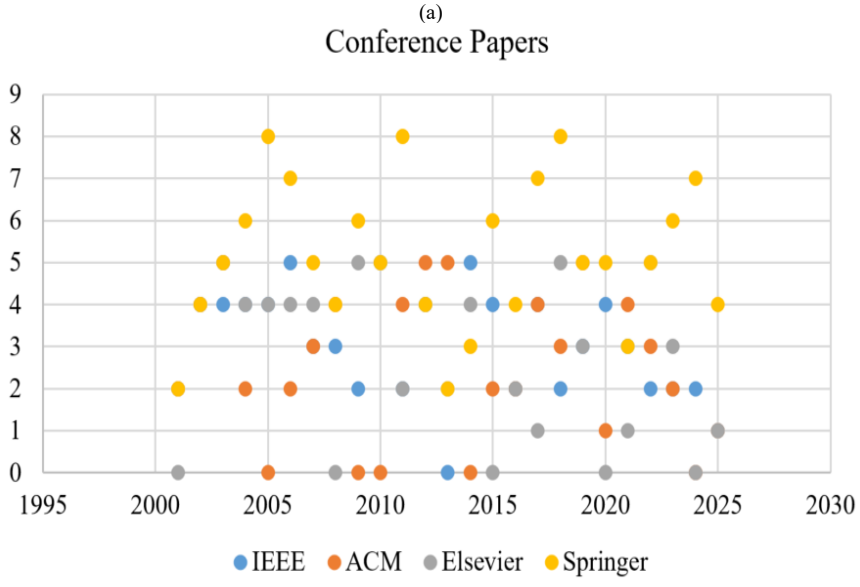
The inclusion and exclusion criteria to select the relevant studies to the research questions are given below in Table II.

TABLE II. THE INCLUSION AND EXCLUSION CRITERIA FOR STUDIES

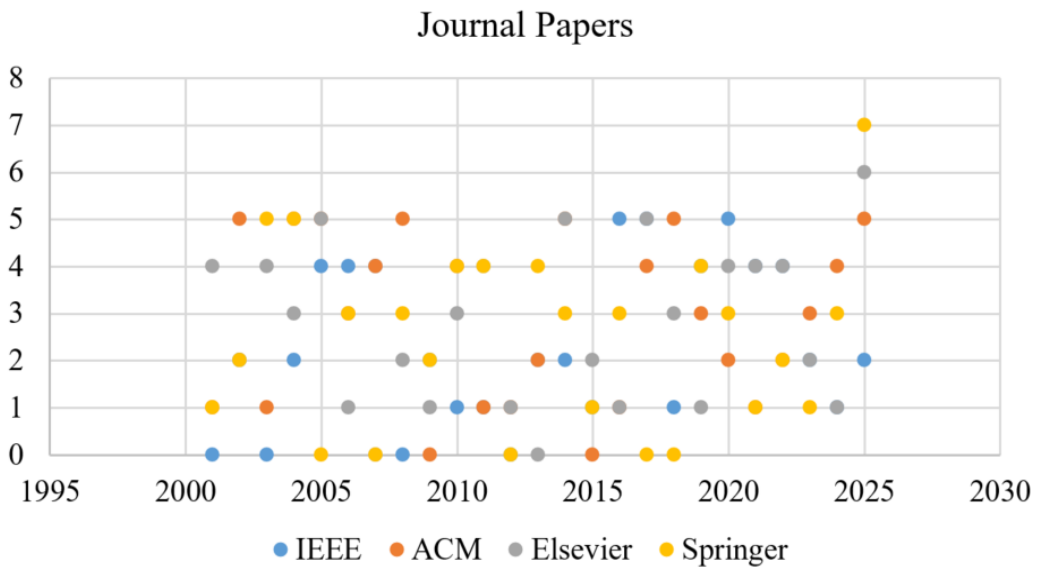
Inclusion Criteria	
IC-1	The paper written in English
IC-2	Peer-reviewed conferences or journals
IC-3	Published after the year 2000
Exclusion Criteria	
EC-1	The paper is not peer-reviewed
EC-2	Papers not written in English
EC-3	Form duplicate studies (one in conference and one in journal on same topic); conference paper was excluded
EC-4	Short research papers (less than 5 pages)



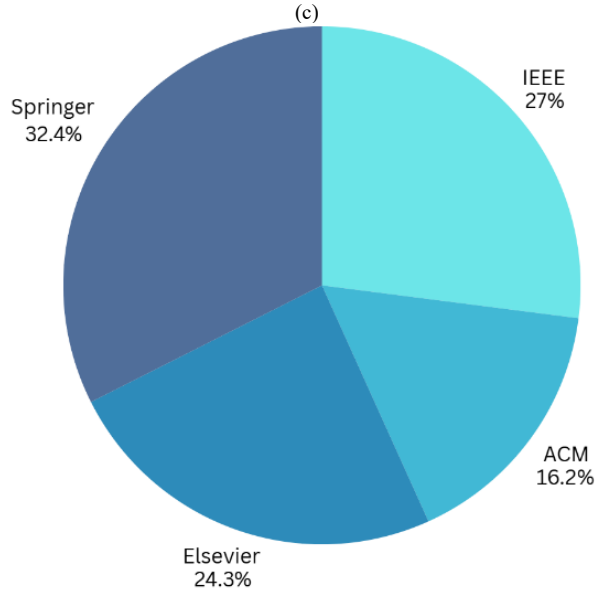
(a)



(b)



(c)



(d)

Fig. 10. (a). Year and publisher-wise study categorization. (b). Year and publisher-wise conference papers. (c). Year and publisher-wise journal papers. (d). Publisher-wise percentage of papers.

1) Selection process: After searching results from the defined databases based on the search strings, the screening process started. At this stage, one of the authors shortlisted papers based on the criteria defined in Table I. At the second stage, full papers were reviewed based on predefined inclusion and exclusion criteria (Table II). In case of confusion, conflict resolution techniques were applied, and the process was reinitiated (Table III). The second stage was concluded with the selection of 77 papers out of 127. After full paper text analysis, another 36 irrelevant papers were dropped. The final 41 papers were assessed for quality against the checklist defined in Table IV, and data were extracted according to Table V. In this survey, forward (Google Scholar) and backward snowballing methods were conducted, resulting in the selection of another VII papers after following all defined selection criteria. In total, this study considers 48 studies, and all data extraction and analysis results are based on these selected studies. The year- and publisher-wise details of selected studies are presented in Fig. 10.

TABLE III. SEARCH AND SELECTION RESULTS

Source	Search Results	Initially scrutinized Results	After full paper analysis	After Snowballing
ACM	17	9	5	6
Springer	25	14	7	9
Elsevier	45	25	11	13
IEEE Xplore	19	12	7	8
Web of Science	21	17	11	12

According to the inclusion and exclusion criteria, 79 papers were rejected among which 31 were not directly related to the business process modeling, 14 were duplicate, 18 were not related to ontology (all were discussing other semantic technologies), 10 were short papers (less than 5 pages) while 6 were basic papers (no details).

2) Selection process problems: During the process of primary studies selection for the SLR, several challenges were faced.

a) One common problem was: the author published the same study at a conference and the extended version in a journal. In this case, the journal paper was considered.

b) Studies with slightly different titles published in multiple journals. In such cases, after reviewing the title, authors, objectives, research problem, and contributions carefully, the whole team discussed and decided to keep or drop these studies.

c) To ensure the validity of the search process. Few research papers were selected based on the prior knowledge and then the most relevant search string was used to verify that string s produce the same results.

To address these challenges, conflict resolution techniques were adopted. Hence, the reliability and completeness of the survey were ensured.

D. Quality Assessment

To assess the quality of selected studies, Quality Assessment (QA) checklists were prepared. The checklist consists of 11 quality assessment criteria used to evaluate each study, as described in Table IV.

TABLE IV. THE QUALITY ASSESSMENT CHECKLIST

Quality assurance Questions	
Q1	Is the study related to business process modeling?
Q2	Are the study objectives SMART?
Q3	Does the study answer any of the research questions?
Q4	Are the research methods clearly described?
Q5	Is the contribution clearly described?
Q6	Is the study evaluation process well explained?
Q7	Is the data along with the source credible?
Q8	Are the implementation methods well explained?
Q9	Are the results and findings properly analyzed and reported?
Q10	Are the results compared with any benchmark?
Q11	Are the research challenges and limitations mentioned?
Q12	Can the results be generalized?

E. Data Extraction

Data extraction from papers is important because it allows researchers to evaluate study quality, along with identifying the strengths and weaknesses. Moreover, upon data extraction, researchers can statistically analyze, summarize, and interpret by ensuring that findings from multiple studies are combined and compared meaningfully. In this regard, data extraction fields used in this study are defined in Table V.

TABLE V. THE DATA EXTRACTION FIELDS

Data Extracted	Description
Research Problem	Research Problem addressed by the study
Research objectives	The research objective of the study
Contributions to existing body of knowledge	The contributions made by the study to the existing body of knowledge.
Data Issues	The challenges faced in data collection
Results	Results reported in the study.
Comparisons results	The results achieved over the benchmark
Key challenges	The main challenges the study encountered

IV. OBPM: A REVIEW

In this section, a concise description of the selected primary studies is presented. Studies are categorized into four major categories (Ontology Driven Process Modeling, Process Integration, & Interoperability, Ontology-Based Business Process Adaptation, and Tools, Frameworks & Applications) based on the research questions. A brief introduction of each category is presented below.

A. Ontology Driven Process Modeling

Ontology-Based Business Process Modeling (OBPM) has emerged as a powerful paradigm for addressing semantic ambiguity, improving interoperability, and enhancing analytical capabilities in modern enterprise systems. Across recent literature, ontologies are consistently positioned as a means to strengthen the semantic foundations of business process management by enabling shared understanding, automated reasoning, and intelligent alignment between business and IT.

In the first paper, Bouroumi et al., [31] discuss how ontologies and context-awareness can improve Business Process Modeling (BPM) by enriching process semantics, supporting context-sensitive adaptations of business processes, bridging gaps between business-level and IT-level understanding, and improving interoperability and reusability of process models. However, the authors have not performed a thorough quantitative evaluation. In another paper, Tao et al., [32] presents a solid contribution in the field of ontology-based information extraction. The authors design an IPO domain ontology capturing key concepts like issuer, underwriter, issue price, risk factors, financial indicators, and company type, etc., from business and financial documents. However, the proposed ontology is primarily used for annotation rather than reasoning or inferencing. A work by Fan et al. [33] proposes a comprehensive process ontology to standardize process elements, clarify terminology, and eliminate inconsistencies. By enforcing explicit semantics, the approach significantly improves communication between stakeholders and prevents misunderstandings in model interpretation. Inter-organizational workflows are addressed in [34]. This paper integrates process mining with ontological mapping to align heterogeneous models across organizational boundaries. This highlights ontology's central role in supporting collaboration where organizations use disparate modeling conventions. Semantic modeling with the help of artificial intelligence is studied in [35]. The authors combine ontological domain knowledge with deep learning to automatically assess process maturity. The hybrid approach represents a growing trend toward augmenting OBPM with

advanced analytics to support intelligent organizational decision-making. Hinkelmann et al. [36] integrate enterprise ontology with an enterprise architecture framework, enabling continuous alignment of business goals with IT structures. Practical alignment is performed by Sánchez et al. [37] using ontology to map processes to web service specifications; the method supports automated model-driven modernization, enabling seamless transition from business design to IT execution environments. BPMO [38] provides a complete formal ontology for describing process models. While Thomas and Fellmann [39] extend this foundation by presenting an ontology-based representation that enhances consistency checking, semantic validation, and process standardization.

All these research works demonstrate that ontology-based approaches significantly improve the clarity, interoperability, intelligence, and performance of business process modeling, positioning OBPM as a critical enabler for next-generation systems.

B. Process Integration and Interoperability

Smith et al. [40] have introduced a logic programming-based framework for business process modeling using ontologies. They used ontologies for reasoning about business process models because traditional process modeling languages like BPMN lack formal semantics, resulting in difficult or incomplete automated reasoning. However, the proposed framework is complex, which makes it difficult to be implemented in large organizations without semantic technology experts. A semantic modeling framework is proposed by Akatkin et al. [41] extends conventional BPM with an ontology. A prototype is implemented for a small-scale government process to show how the framework has improved interoperability. The paper should discuss how ontologies will be maintained, versioned, and aligned with regulatory changes. In a paper [42], the author has merged the process workflows with semantic domain knowledge for improving interoperability, model consistency, and intelligent system behavior. Ontologies developed in isolation lead to inconsistencies, limited interoperability, and difficulties in reusing knowledge across systems. Pfaff et al. [43] have designed an ontology-based framework to support semantic data integration within IT benchmarking environments. Authors argue that different organizations generate different type of data which might create significant challenges like data heterogeneity, inconsistency, and incompatibility. However, for evaluation, a limited dataset is used. The authors have not discussed the integration tools, automated reasoning, or user interface considerations, which makes their work difficult to implement by non-technical users.

The above-discussed papers highlight the advancement of Semantic Business Process Management and ontology-based integration. The papers collectively discuss the strengths of formal semantics for process modeling and improving business integration through shared ontologies. Whereas the proposed works are facing challenges of scalability, ontology maintenance, and deployment at the enterprise level.

C. Ontology-Based Business Process Adaptation

In this section, different research works are reviewed which collectively highlight how ontology engineering enhances business process modeling, semantic consistency, and adaptability.

In a paper, Preisig et al. [44] have proposed an ontology to capture physical entities formally along with their relationships, constraints, and operational dependencies. In their work, they demonstrated an example of linking physical topology with business processes and performed automated reasoning to check consistency of system structures and detect invalid process sequences. However, they did not discuss the scalability issues for large-scale physical systems. In a paper, Kord et al. [45] focused on dynamic business process adaptation through context ontology. To demonstrate their achievement, a context ontology along with reasoning rules was used to infer the need of adaptation. However, they have not discussed how their proposed system can be applied in a real-world industrial scenario. Samhan et al. [46] have extended the Riva methodology [52] for business process architecture. They extended the Riva methodology with an ontology for automatic interpretation and converted an informal architectural methodology into a semantically enriched framework. But the proposed system does not provide support for executable process modeling tools. Poss et al. [47] in their work have focused on the operational aspects (tasks, resources, preconditions, and effects) of business process modeling. They claim that they have provided a foundation for process execution verification, simulation, and compliance enforcement. However, the proposed ontology is very complicated, which might confuse typical process designers. Thuan et al. [48] have addressed the process flexibility challenge of business process modeling for organizations that work in highly dynamic environments. However, their work highly relies on manually encoded rules. Moreover, the performance issues of their work are also not fully assessed.

All above discussed papers demonstrate that traditional BPM languages lack semantic clarity, leading to ambiguity and inconsistency. Whereas ontologies support precision, adaptation, and enable reasoning-based improvements in flexibility. However, ontology adoption in terms of complexity and scalability is a big challenge for the next-generation intelligent BPM systems.

D. Tools, Frameworks & Applications

Ontology-based techniques, due to formal semantics, reasoning, and improved interoperability across heterogeneous business environments, are becoming fundamental in Business Process Management. In this section, the literature on Tools, Frameworks & Applications is reviewed for synthesizing their contributions, methodological approaches, limitations, and collective significance for the field of semantic business process management.

Ghasemi et al. [49] have developed a unified ontology to classify and configure the business process management techniques which consists of Six Sigma, Lean, BPMN, simulation tools, and workflow engines etc. The major contribution of this work is the BPM maturity assessment framework, which can map organizational needs to appropriate techniques via semantic reasoning. However, the proposed mechanism is not evaluated through different case studies for performance. In a paper, Annane et al., [50] evaluate different

business process ontologies (on criteria like expressiveness, modularity, reasoning capabilities, domain flexibility, and alignment with BPMN) to know their suitability for task monitoring. Through evaluation, they identified ontological gaps like limited support for adaptation, temporal constraints, and the organizational context. However, it is a conceptual study that must be supported with empirical evaluation using real-time monitoring systems. Adams et al., [51] have addressed the challenge of complexity of large business process models through ontology-based views which represent different dimensions of a process like functional, behavioral, organizational, and data perspectives. The proposed mechanism has the ability to generate automatically customized views based on stakeholder roles and analytical needs without losing semantic consistency. However, the process of designing these views requires a considerable semantic expertise.

In the above discussed papers, it is revealed that ontologies enable precise definitions of process models by eliminating ambiguities inherent in traditional modeling languages and the formal semantics enables the automated process validation, compliance checking, and real-time decision support. Moreover, ontology-based process modeling approaches improve context-awareness and interpretability in discovering and analyzing business processes. However, certain challenges are still open to investigation. For example, ontology creation & maintenance and computational overhead along with scalability issues are not studied in detail considering industry-grade case studies.

E. Comparative Evaluation

In this section, a comparative evaluation of ontology-based business process modeling studies is performed on criteria: Tool Support, Empirical Validation, Quantitative Metrics, Reasoning Capabilities, Adaptability, Context-awareness, Generalizability, and Limitations. Criteria requirements are divided into two sections 1) general requirements (Tool Support, Empirical Validation, Quantitative Metrics) and 2) ontology specific requirements (Reasoning Capabilities, Adaptability, Context-awareness, Expressiveness, and Limitations). The objective of the evaluation is to know the strengths and limitations of these studies. Evaluation criteria requirements are defined in Table VI, while the evaluation results are presented in Table VII and Table VIII, respectively.

TABLE VI. LITERATURE EVALUATION CRITERIA REQUIREMENTS

Evaluation criteria	Description
Tool Support	Availability of software tools for BPM platforms or support for automation and semantic validation
Empirical Validation	Experimental demonstration to improve accuracy or efficiency of ontology or non-ontology BPM approaches
Quantitative Metrics	Performance measures like modeling time or complexity analysis of ontology integration
Reasoning Capabilities	Incorporation of formal semantics to demonstrate automated process validation, compliance checking, and real-time decision support.
Adaptability	System has the ability to modify process execution based on semantic or contextual conditions.
Context-awareness	Research captures dynamic context like roles, locations, events, and rules.
Generalizability	Flexibility to adapt to evolving processes and ontologies or have potential for reuse in other knowledge-intensive domains
Limitations	The constraints (methodological, theoretical, or practical) that restrict the findings, scope, or generalizability.

TABLE VII. STUDIES EVALUATION ON GENERAL CRITERIA REQUIREMENTS

Study	Tool Support	Empirical Validation	Quantitative Metrics
[31]	Prototype-level	Small case examples	Not reported
[32]	Prototype implemented for document analysis	Experiments on IPO documents	Extraction accuracy / precision/recall
[33]	Prototype and mapping examples discussed	Experimental evaluation	Some evaluation metrics (e.g., ambiguity reduction measures).
[34]	Methodology and implementation details hinted	Experiments / case studies	Process discovery / conformance metrics reported.
[35]	Prototype and experimental pipeline described.	Datasets used for capability assessment.	Performance metrics for assessment (accuracy, F1, etc.).
[36]	Methodological framework; tool prototypes in some implementations	Case studies in industry contexts	Not the main focus; some alignment indicators discussed.
[37]	Prototype tools and model transformations	Demonstration and experiments reported.	Some performance / correctness metrics for transformations
[38]	Research prototypes and semantic tooling described	Example scenarios and prototype evaluations.	Not central; mostly conceptual/qualitative evaluation
[39]	Implementation details and prototype frameworks	Examples and small evaluations.	Not emphasized
[40]	Implementation (logic programming engine).	Examples and reasoning case studies provided	Some performance discussions; not heavily empirical.
[41]	Frameworks and prototypes discussed in paper	Government examples; conceptual demonstrations	Not the focus
[42]	Methodology and tooling recommendations, possible prototype	Case studies	No central methodological focus
[43]	Web-based architecture for ontology-driven integration	Demonstrations and validation examples	Integration performance / mapping accuracy discussed.
[44]	Implementations and examples in journal article.	Case examples and model demonstrations	Not emphasized
[45]	Prototype and reasoning components described	Examples and limited experiments	Not heavily emphasized
[46]	Conference paper with prototype	Limited empirical evaluation	Not reported

[47]	GitHub implementation available	Prototypical demos	Not a major focus; qualitative validation present.
[48]	Conceptual prototype; limited tooling.	Conceptual validation and examples	Not central; few evaluation metrics.
[49]	Ontology artifact published	Limited; mainly conceptual and illustrative.	Not emphasized.
[50]	Comparative analysis; no single tool proposed	Analytical comparison	Comparative criteria used; not performance metrics.
[51]	Prototype and formal definitions	Examples and comparative discussion	Not central; usability discussion present.

TABLE VIII. STUDIES EVALUATION ON ONTOLOGY SPECIFIC CRITERIA REQUIREMENTS

Study	Reasoning Capabilities	Adaptability	Context-awareness	Generalizability	Limitations
[31]	Ontology reasoning for context interpretation (limited).	Targets adaptive behaviors tied to context	Central focus; explicit context modeling	Domain-agnostic idea; validation limited	Evaluation small-scale; maintenance overhead noted
[32]	Semantic enrichment and mapping; basic inference for IE.	Focused to IPO domain; adaptable with new ontology work.	No central document	Needs rework for non-financial domains	Domain specificity; ontology construction cost.
[33]	Ontology reasoning for equivalence	Aimed at heterogeneous modeling environments	Not in focus	Intended to be transferable across domains	Ontology engineering overhead; scalability concerns for very large mode
[34]	Semantic mapping to align logs	Built for heterogeneous partner systems.	Indirect focuses on organizational context.	Targeted to inter-org settings; transferable with mapping	Complexity with noisy or incomplete logs; mapping effort required.
[35]	Ontology supports semantic features; DL handles pattern detection	Relatively adaptable; dependent on retraining and ontology updates	Limited explicit context modeling	Promising; requires dataset and ontology alignment for new domains	Data requirements for DL; ontology upkeep.
[36]	Ontology-based reasoning for alignment checks	Designed for organizational adoption; flexible modelling approach.	Captures organizational and technical context.	Strong enterprise applicability; requires customization	Implementation cost; governance/maintenance challenges
[37]	Semantic mapping and matching for service binding.	Focus on re-use in legacy modernization scenarios	Limited; more about mapping semantics to services	Applicable to modernization projects broadly	Complexity of mapping and runtime orchestration.
[38]	Supports reasoning for discovery, validation and mapping.	Designed to be extensible and reusable.	Includes organizational and service context modelling	Developed to be reusable across domains	Ontology complexity need tool support.
[39]	Used for validation and semantic queries.	Focused on improving model interchange and reuse.	Some attention to context and organizational info.	Designed for general BPM settings.	Tool maturity and scaling concerns
[40]	Strong (logical inference, consistency checks).	Method applicable where formal verification needed.	Not primary	Technically general but requires expertise	Complexity; scalability for very large models
[41]	Supports semantic alignment for cross-agency workflows	Tailored to public services; generalizable with work	Includes citizen/service context in examples	Framework transferable to other public sectors.	Organizational barriers; real deployments are complex
[42]	Enables semantic checks during development	Designed to fit organizational practices	Can incorporate context if modeled	Method intended to be generic	Resource intensity; requires ontology expertise.
[43]	Semantic mapping and mediation supported	Domain-specific but extensible	Limited to IT benchmarking context.	Useful template for other benchmarking domains	Domain restriction; ontology maintenance needs
[44]	Ontology used for semantics and consistency checks.	Targeted at engineering/process industries	Physical context modeled explicitly.	Good for industrial contexts	Domain specificity; complexity of physical modelling.
[45]	Central: ontology reasoning for adaptation decisions	High; designed for dynamic adaptation	Core focus.	Applicable where contextual data is available	Real-time data acquisition and reasoning overhead
[46]	Basic semantic alignment and mapping	Focused to RIVA architectures may be generalize	Some architectural context captured	Needs adaptation for other architectures	Specialized scope; limited empirical proof.
[47]	Reasoning for resource allocation and integration checks	Targeted to SMEs but extensible	Includes operational context (equipment state).	Applicable to domains with physical resources.	Implementation maturity and runtime constraints.
[48]	Enables classification and reasoning about flexibility types.	Helps design flexible processes; supports reasoning about variability.	Flexibility linked to context and exceptions.	Intended to be broadly applicable.	Needs tooling and empirical validation
[49]	Supports semantic queries and mappings	Useful for tool recommendation tasks; needs updating.	Not central.	Broad for BPM community; needs maintenance.	Rapid tool evolution requires updates; limited validation.

[50]	Assesses ontologies' support for monitoring reasoning.	Guides selection/extension for monitoring tasks	Evaluates how ontologies capture execution context.	Comparative results help across domains.	Lacks experimental runtime evaluation.
[51]	Ensures semantic consistency across views.	Highly adaptable by adding view extensions.	Views can encode contextual perspectives.	Broad applicability to enterprise modeling.	Complexity in authoring view-extensions; tooling required.

TABLE VIII (A). STUDIES EVALUATION ON ONTOLOGY SPECIFIC CRITERIA REQUIREMENTS

Paper	Ontology Complexity	Scalability & Performance	Tool Integration	Heterogeneity	Semantic Interoperability	Syntactic Interoperability	Process Interoperability	Context Interoperability
[31]	✓	✓	✓		✔		✔	✔
[32]	✓	✓		✗	✔	✔		
[33]	✓		✓		✔		✔	
[34]				✓		✔	✔	
[35]	✓	✓			✔			
[36]	✓		✓		✔	✔	✔	
[37]	✓		✓		✔	✔	✔	
[38]	✓		✓	✗	✔		✔	
[39]	✓	✓	✓		✔		✔	
[40]	✓	✓		✗	✔		✔	✔
[41]	✓	✓	✓	✓	✔	✔	✔	
[42]	✓		✓		✔	✔	✔	✔
[43]	✓			✓	✔	✔		
[44]	✓	✓			✔		✔	
[45]	✓	✓			✔		✔	✔
[46]	✓		✓	✗	✔		✔	
[47]	✓	✓	✓	✗	✔	✔	✔	
[48]	✓	✓			✔		✔	✔
[49]	✓			✓	✔			
[50]	✓		✓	✓	✔		✔	
[51]	✓	✓	✓					

The above-evaluated papers describing frameworks, tools, and methodologies provide thorough and clear but varying explanation details. Papers focusing on ontology-based business process modeling [38, 40] provide a strong semantic foundation by clearly differentiating the process concepts, control-flow structures, behavioral semantics, and organizational perspectives. Whereas few papers [35, 36] explain how ontologies improve knowledge representation, data interoperability, and process alignment, they combine ontology with deep learning or enterprise architecture and prefer architectural descriptions over detailed ontology specifications. Papers addressing context-aware modeling [31, 45] provide detailed descriptions of contextual elements like environmental variables, dynamic conditions, roles, and resources. Although these elements strengthen the conceptual clarity of context modeling but also increase the complexity of the model.

However, a number of framework-based papers are conceptual, discussing architectural designs without presenting any software artifacts. For example, papers focusing on process capability assessment [35] or enterprise alignment [36] describe tool requirements but do not produce a complete toolset. In short, about thirty-three percent of papers offer strong tool support, while thirty-three percent of papers provide partial or prototype-level tools, and the remaining are just conceptual. Papers providing semantic model-based prototypes, especially on the context adaptation [45] and semantic process alignment [37] provide working implementations by integrating workflow engines with semantic annotation systems.

The evaluated works present strong development in ontology-based business process modeling while considering enhanced semantic expressiveness, improved reasoning capabilities, better interoperability, enriched context modeling, improved semantic clarity and alignment, and the combination of ontologies with process mining using ML. Whereas, the major weaknesses of these studies are limited empirical validation, established tool support, the use of highly varying quantitative metrics, and evaluation (often limited to case studies), restricts the generalizability.

V. RESULTS AND DISCUSSION

Most of the papers cover a broad range of BPM lifecycle phases like process modeling-focused works [38, 51] emphasize

semantic representation and structural modeling; context-aware approaches [31, 45, 60] extend models to dynamic adaptation; papers on process integration [34, 37, 46] highlight cross-enterprise interactions; process flexibility paper [48] and ambiguity resolution paper [33] focus on behavioral aspects. Whereas a paper considering domain-specific ontology integration [43] constrains the scope to IT benchmarking. Papers in the process mining and integration [34, 50] domain have a tendency to consider robust empirical investigations using event logs, organizational datasets, or inter-organizational workflows. These studies evaluate performance, model accuracy, or integration improvements based on real-world or semi-real datasets. While papers on ontology engineering [38, 39, 51] include case studies involving business process scenarios in domains like manufacturing, e-government, logistics, and service management. These case studies help to illustrate the ontology applications but lack quantitative or statistically significant validation. Whereas papers that have considered ontology with AI [35] (hybrid deep learning + ontology) consider detailed evaluation methods but primarily focus on model performance rather than ontology quality. Likewise, some papers lack empirical validation altogether, particularly high-level conceptual works on enterprise architecture alignment [36] or adaptability frameworks [48]. These limitations reduce practical reliability and impact. Traditional ontology-based modeling papers [38, 39] treat processes as static artifacts. Integration works [37, 41] occasionally incorporate adaptation but typically focus more on interoperability than real-time flexibility. However, despite these limitations, the papers collectively advance ontology-based BPM significantly.

A. RQ1: What Challenges or Limitations are Reported in Ontology-Based BPM Research?

Ontology-based Business Process Management (OBPM) has been widely studied for its potential to enhance semantic clarity, interoperability, process automation, and organizational alignment. However, the major limitation reported is the complexity involved in developing and maintaining ontologies [31, 38, 39]. Creating semantically rich process ontologies depends on domain knowledge and expertise in ontology engineering. Many organizations lack personnel with such skills, making ontology construction costly and time-consuming. Moreover, a few common challenges discussed in each study are the narrowness of ontological scope, absence of formal reasoning, limited empirical evaluation, lack of standardized tool support, and integration issues [37, 47].

Many studies also highlight scalability and performance issues [33, 45, 51]. Reasoning engines used for ontology alignment, context inference, or semantic disambiguation face such issues in large organizations. Kord et al., [45] discuss that in real-time environment, the reasoning can be computationally costly due to the fast change in context or in data, resulting in delays in monitoring and decision support [51].

Another reported challenge is the semantic heterogeneity across organizations. A number of researchers [33, 41, 43] have noted that different organizations use different ontological terms, modeling conventions, and process structures, which requires a lot of effort for consensus building and mapping. A few studies have pointed out the need for limited real-world validation [46, 48]. In [48], a rich conceptual model is presented that lacks extensive empirical testing across industries. As a result, scalability, user adoption, and practical impact are not evaluated.

Further research dimensions reported in [32] and [35] are the integration of ontologies with advanced analytics and machine learning. As ontologies provide structure, while machine learning systems demand large datasets and flexible representations. Hence, integration of both techniques may result in clashes between symbolic and statistical paradigms. Similarly, studies like [50] and [49] report the lack of standardization across BPM ontologies. That results in duplication of concepts, incompatibilities, and difficulties in selecting the right ontology for a given use case.

B. RQ2: What Ontological Types are Required to Represent BPM Effectively?

Ontology-based Business Process Management (OBPM) research has emphasized a set of core ontological types required to represent business processes in a semantically rich, interoperable, and machine-processable way. These types can broadly be categorized into process structure, behavioral semantics, organizational context, data and information flow, resources and capabilities, and finally into contextual or environmental factors. Each category has a strong representation of processes that enable reasoning, interoperability, and automation.

Similarly, few ontology-based BPM studies [33, 38, 39, 50, 51] represent the basic building blocks of a process like activities, events, control-flow constructs, composite activities, and the sequence relations. These building blocks formalize the BPM structure and eliminate semantic ambiguity by defining the meaning of each element in a machine-processable format, and without these structural types, process reasoning and consistency checking become impossible [39, 41]. A few papers [37, 40, 47] discuss that structural elements alone are insufficient. For effective business process modeling, demand ontologies to capture execution semantics, including pre and post-conditions, state transitions, temporal constraints, concurrency and synchronization semantics, along with the business rules. These types allow ontologies to support logical reasoning, detect inconsistencies, and enable automated process execution.

Papers [36, 41, 46] have discussed the impact of the integration of context into BPM ontologies. Organizational context includes actors, roles, management units, actor responsibilities, and the relationships among stakeholders. This context helps in mapping the business processes to the enterprise architecture, ensuring alignment between business and IT. In addition, several papers [35, 37, 49] highlight that BPM ontologies must represent capabilities and resources to support dynamic adaptation and process capability assessment. This includes human resources, system applications, services, skills, and competencies. It allows process execution engines to bind tasks to appropriate real-world resources. Papers [32, 43, 49] emphasize the need to integrate semantic data models like data objects or business entities, information flows, document types, and Input, output dependencies into BPM ontology frameworks.

These data models enable semantic data integration, reduce ambiguity, and support analytics and process mining. Furthermore, papers [31, 45, 48, 44] highlight the need to represent context (environmental, organizational, user, and device context) for adaptive and flexible processes. These types of contexts are critical for context-driven process adaptation and dynamic reconfiguration.

The reviewed literature on BPM ontology, effective representation requires a multi-layer ontology that integrates structural, behavioral, organizational, resource, data, and contextual types. These ontological categories enable semantic clarity, automated reasoning, interoperability, and adaptability, which is the ultimate goal of present ontology-based BPM research.

C. RQ3: How can OBPM Support Interoperability Between Heterogeneous Business Organizations?

Ontology-Based Business Process Management plays a central role in enabling interoperability between heterogeneous business organizations by providing a shared semantic foundation for representing, aligning, and integrating processes. A number of papers [32, 33, 38, 39, 43, 50] contribute to interoperability through four major mechanisms: semantic harmonization, cross-organizational alignment, context-aware integration, and machine-reasonable process representations. Together, these mechanisms allow organizations with different models, terminologies, systems, and workflows to communicate, integrate, and execute processes coherently. OBPM has the ability to harmonize semantics across organizations that use different vocabularies or modeling conventions. Authors of papers [33, 38, 50] show that ontologies remove ambiguity from process elements by providing precise definitions of tasks, events, roles, resources, and data objects. When organizations map to a common ontology, they can compare and integrate processes that were previously incompatible due to terminology differences.

Papers [32, 43] report on how ontology-based semantic data integration consolidates heterogeneous information sources, which is crucial when organizations exchange documents or transactional data. Similarly, a few other studies [34, 36, 37, 41] demonstrate that OBPM supports interoperability by aligning business processes across organizational boundaries. Ontologies can map one organization's processes to another through process mapping and equivalence identification, cross-organizational workflow integration, semantic matching between business services, and automatic transformation of process models into executable forms. Aouachria et al. [34] have demonstrated how ontology-driven process mining can detect correspondences between heterogeneous inter-organizational processes, enabling integrated workflow execution. Akatkin et al., [41] discusses how shared process ontologies facilitate the integration of seamless public services in an e-government scenario. Interoperability among organizations usually fails when processes behave differently in different contexts. OBPM uses context ontologies to capture environmental conditions, policies, and constraints, etc. [31, 45, 48]. Contextual knowledge can be used in business process models to adapt processes when exchanged in organizational environments or to ensure that cross-organizational workflows respect local rules and constraints and, finally, to resolve conflicts caused by differing operational contexts. In the case of the reasoning capability of OBPM, papers [40, 47, 42] show that ontology-based models support automated reasoning to check compatibility, identify inconsistencies, and verify the alignment of processes across organizations by allowing the validation of the process conformance, resolution of semantic conflicts, inferencing of the missing relationships, and composition or orchestration of services [37].

Papers on OBPM enable interoperability by providing a shared vocabulary, a formal mechanism for alignment, adaptive contextualization, and automated reasoning. It allows organizations to collaborate effectively, integrate seamlessly, and automate cross-organizational processes without worrying about the differences in terminology, technology, or workflow design. OBPM, therefore, can resolve the issue of interoperability among organizational business process integrations.

D. RQ4: How have Ontology-Based BPM Approaches been Evaluated or Validated in Existing Studies?

In ontology-based BPM research, evaluation and validation methods vary with respect to semantic enrichment, process integration, context modeling, system alignment, and interoperability. These methods include case studies, prototype implementation, comparative analysis, reasoning tests, process mining, and expert judgment [31, 51]. In the case of study evaluations, many studies [31, 32, 35, 41] are validated through domain-specific ontologies demonstrating applicability in real-world process scenarios. In these studies, context-based modeling is validated through contextualized business processes to show increased expressiveness and accuracy. For IPO documents analysis, financial prospectus datasets are used to validate information extraction accuracy [32]. For process capability assessment, deep learning and ontology reasoning are integrated [35] to test in organizational scenarios. Papers [38, 39, 45, 47] introducing new modeling techniques are often supported with tools, like in a paper proposing Business Process Modeling Ontology [38] is evaluated with a prototype demonstrating how models can be instantiated and queried. In the semantic process modeling prototype [39], the author demonstrates model creation, storage, and reasoning in an ontology-backed environment. Similarly, in a paper on process adaptation systems [45], the author validated context-driven adaptation via prototype reasoning engines. Likewise, the author of operational perspective formalization [47] has performed a tool integration experiment to validate interoperability.

Since formal semantics are central to OBPM, many studies validate their ontologies through logic-based reasoning tests. For example, Smith et al. [40] used logic programming to test completeness, consistency, and inferential accuracy. In another study [33], the author evaluated semantic disambiguation by checking whether the ontology correctly resolves conceptual overlaps. Similarly, Thuan et al. [48] tested flexible process configurations using reasoning rules to validate variability and adaptability. Furthermore, few studies compare new ontologies with existing standards, frameworks, or business process ontologies. Annane et al. [50] evaluated multiple ontologies on expressiveness, support for monitoring constructs, and

reasoning capabilities. Aouachria et al. [34] used process mining to validate semantic alignment by showing improved cross-organization trace matching. In another study, Pfaff et al. [43] evaluated effectiveness through data integration tests across heterogeneous IT performance datasets.

However, in the case of validation through expert judgment to assess ontology structure, coverage, and practicality. Hinkelmann et al. [36] validated Enterprise ontology alignment through expert walkthroughs. For the validation of Riva-driven architecture adaptation [46], the author applied expert assessment to validate the generality of adapted ontological constructs. OBPM approaches proposed in the above-discussed studies are validated through case studies, prototypes, logical reasoning, comparative analysis, process mining, data integration experiments, and expert evaluation, reflecting the multidimensional nature of ontology-driven BPM research.

In Table IX, a few prominent review studies are presented that discuss semantic business process management and ontology-based management systems. Unlike these reviews [55 - 57], this study comprehensively reviews the ontology-based business process modeling approaches, identifies emerging research trends, analyzes challenges and opportunities, and proposes a conceptual Input-Process-Output (IPO) framework derived from the reviewed literature.

TABLE IX. EXISTING SYSTEMATIC REVIEWS ON BPM

Review	Focus	Number of Studies	Research Gaps
[53]	Semantic technologies to business process management (SBPM)	36	Limitations in semantic BPM (SBPM)
[54]	Integration of Knowledge Management and business process management	88	Challenges in integrating BPM and KM
This Review	Ontology-Based BPM	49	Comprehensive taxonomy and trend analysis of semantic BPM

VI. CONCLUSION AND FURTHER WORK

Business Process Modeling (BPM) represents an organization's business processes in a formal and structured way. A business process consists of a set of interrelated activities, their dependencies, roles, data flows, and constraints that collectively achieve a specific organizational goal, such as order fulfillment or customer support. Stakeholders, analysts, and IT professionals use a few common modeling notations of BPM, like BPMN, UML, Petri Nets, and EPCs for a shared understanding of process analysis, compliance verification, performance measurement, redesign, and automation. However, traditional BPM approaches lack explicit semantics, which leads to several challenges, like business process models frequently suffering from ambiguous interpretations of activities (same concept may refer to different meanings in different contexts), roles, and data elements, resulting in limited automated analysis and restricted communication among stakeholders. Similarly, different tools and languages used in modeling cannot interoperate seamlessly. This is a big problem in dynamic environments where multiple systems must collaborate. Moreover, the difference between high-level business process models and their technical implementation may lead to increased maintenance costs and inconsistent behavior of business systems. Additionally, BPM notations focus on graphical representation and provide limited support for automated reasoning, validation, and consistency checking, resulting in a restricted ability of error detection and optimization of processes systematically. However, in order to resolve these issues, ontologies can be used because ontologies provide a formal, explicit, and shared conceptualization of a domain by defining concepts, relationships, constraints, and rules in a machine-interpretable manner. Therefore, supporting process models with ontologies can ensure semantic consistency across models, resulting in reduced ambiguity and misinterpretation. Although many researchers have worked on ontology-based business process modeling, a lot of issues need to be investigated. Ontology-based BPM frameworks proposed by many studies still lack the adoption of standards, causing interoperability issues. So, developing domain-independent ontologies with the ability to be adapted to specific organizational contexts while maintaining semantic consistency across systems is a challenge. Similarly, scalability and performance optimization issues need to be addressed in systems where ontologies are used with machine learning (integration addresses the issue of computationally expensive ontology-based reasoning in large-scale or data-intensive environments). The integration of both techniques may result in clashes between symbolic and statistical paradigms. Moreover, the integration of ontologies (for improving transparency, trust, and interpretability) with deep learning, process mining, and decision support with explainable AI in BPM is a highly promising research direction. Within the OBPM body of knowledge, two important aspects are crucial. First, how should the process ontology be designed? Second, how should the designed process ontology be used or leveraged? Further research needs to explore the best approach to utilize the ontology supporting business process modeling, which will lead to a fundamental contribution to OBPM literature.

A. Review Limitations

This review has a number of limitations. 1) due to English language publications consideration only, a large number of relevant studies published in other languages are omitted resulting in compromised survey quality. 2) Grey literature, industrial reports, and technical white papers are not considered, which is another quality question on the survey. 3) Year-wise restricted publication selection may have omitted relevant studies. 4) other databases must also be considered as specified above to enhance the likelihood of relevant studies consideration. Overall, the incorporation of these limitations in the literature may provide a more comprehensive perspective on the selected topic. Moreover, as this study is descriptive in nature, a formal statistical sensitivity analysis is not performed.

REFERENCES

[1] Moreira, S. A. S., & Dallavalle, S. (2024). Unlocking the potential of digital innovations on BPM: a bibliometric and systematic literature review. Business Process Management Journal.

[2] Bitkowska, A., Detyna, B., & Detyna, J. (2023). Towards integration of business process management and knowledge management. IT systems' perspective.

[3] Koblianska, I., Varakin, D., Pihul, O., Somushkin, V., & Glukh, V. (2023). Review of scientific literature on BPM concept in social sciences. Problems and Perspectives in Management, 21(3), 84.

[4] Medina Garcia, S., Medina-Marin, J., Montaño-Arango, O., Gonzalez-Hernandez, M., & Hernandez-Gress, E. S. (2023). A Petri net approach for business process modeling and simulation. Applied Sciences, 13(20), 11192.

[5] Qin, J.; Zhao, N.; Xie, Z.; Mo, Q. Business Process Analysis Method Based on Petri Nets. In Proceedings of the 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 21–23 July 2017; pp. 218–223

[6] Indulska, J. Recker, M. Rosemann, P. Green, Business process modeling: current issues and future challenges, Advanced Information Systems Engineering, Springer 2009, pp. 501–514.

[7] Trachenko, I., Lazorenko, I., Maslennikov, Y., Hrinchenko, Y., Arsawan, I., & Koval, V. (2021). Optimization modeling of business processes of engineering service enterprises in the national economy. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, 2021(4), 165-171.

[8] Mili, G. Tremblay, G.B. Jaoude, E. Lefebvre, L. Elabed, G.E. Boussaidi, Business process modeling languages: sorting through the alphabet soup, ACM Comput. Surv. (43:1) (2010) 4–56.

[9] Pufahl, L., Zerbato, F., Weber, B., & Weber, I. (2022). BPMN in healthcare: Challenges and best practices. Information Systems, 107, 102013.

[10] Polančič, G., & Cegnar, B. (2017). Complexity metrics for process models–A systematic literature review. Computer Standards & Interfaces, 51, 104-117.

[11] Ayech, H. B. H., Ghannouchi, S. A., & Amor, E. A. E. H. (2021). Extension of the BPM lifecycle to promote the maintainability of BPMN models. Procedia Computer Science, 181, 852-860.

[12] Ireddy, A. T., & Kovalchuk, S. V. (2023). An experimental outlook on quality metrics for process modelling: a systematic review and meta analysis. Algorithms, 16(6), 295.

[13] Zhou, C., Zhang, D., Chen, D., & Liu, C. (2023). Business process complexity measurement: A systematic literature review. Ieee Access, 11, 47940-47955.

[14] Recker, M. Rosemann, M. Indulska, P. Green, Business process modeling—a comparative analysis, J. Assoc. Inf. Syst. April 24, 2009 (10:4) (2009) 333–363.

[15] Van der Aalst, W. M. (2013). Business process management: a comprehensive survey. International Scholarly Research Notices, 2013(1), 507984.

[16] Sahli, A., Mejri, A., & Ghannouchi, S. A. (2025). Leveraging Process Mining and Machine Learning to Enhance Performance Analysis.

[17] Fattouch, N., Lahmar, I. B., & Boukadi, K. (2024). A model-driven approach for IoRT-aware business process engineering. International Journal of Computer Information Systems and Industrial Management Applications, 16(3), 19-19.

[18] Von Rosing, M., White, S., Cummins, F., & De Man, H. (2015). Business Process Model and Notation-BPMN.

[19] Amjad, A., Azam, F., Anwar, M. W., Butt, W. H., & Rashid, M. (2018). Event-driven process chain for modeling and verification of business requirements–a systematic literature review. Ieee Access, 6, 9027-9048.

[20] Issahaku, Fl.Y., Lu, K., Xianwen, F. et al. An overview of semantic-based process mining techniques: trends and future directions. Knowl Inf Syst 66, 5783–5827 (2024). https://doi.org/10.1007/s10115-024-02147-x

[21] Sweller, J. (2011). Cognitive load theory. In Psychology of learning and motivation (Vol. 55, pp. 37-76). Academic Press.

[22] Barbara A. Kitchenham. 2006. Evidence-based software engineering and systematic literature reviews. In In Proceedings of the 7th International Conference on Product-Focused Software Process Improvement, PROFES 2006, Amsterdam, The Netherlands, June 12-14, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 4034). 3.

[23] Weske, M. (2012). Business process management architectures. In Business process management: Concepts, languages, architectures (pp. 333-371). Berlin, Heidelberg: Springer Berlin Heidelberg.

[24] Szelągowski, M., & Lupeikiene, A. (2020). Business process management systems: evolution and development trends. Informatica, 31(3), 579-595.

[25] Aalst van der, W. M. (2004). Business process management: a personal view. Business Process Management Journal, 10(2).

[26] Van Der Aalst, W. M., La Rosa, M., & Santoro, F. M. (2016). Business process management: Don't forget to improve the process!. Business & Information Systems Engineering, 58(1), 1-6.

[27] Recker, J., Lukyanenko, R., Jabbari, M., Samuel, B. M., & Castellanos, A. (2021). From representation to mediation: a new agenda for conceptual modeling research in a digital world. MIS quarterly, 45(1), 269-300.

[28] Kir, H., & Erdogan, N. (2021). A knowledge-intensive adaptive business process management framework. Information Systems, 95, 101639.

[29] Silva, V. T., Dos Santos, J. S., Thiago, R., Soares, E., & Azevedo, L. G. (2022). OWL ontology evolution: understanding and unifying the complex changes. The Knowledge Engineering

[30] Corea, C., Fellmann, M., & Delfmann, P. (2021, October). Ontology-based process modelling-will we live to see it?. In International Conference on Conceptual Modeling (pp. 36-46). Cham: Springer International Publishing.

[31] Bouroumi, J., GuermaH, H., & Nassar, M. (2021). Enhancing business process modeling with context and ontology. International Journal of Advanced Computer Science and Applications, 12(9).

[32] Tao, J., Deokar, A. V., & El-Gayar, O. F. (2014, January). An ontology-based information extraction (OBIE) framework for analyzing initial public offering (IPO) prospectus. In 2014 47th Hawaii International Conference on System Sciences (pp. 769-778). IEEE.

[33] Fan, S., Hua, Z., Storey, V. C., & Zhao, J. L. (2016). A process ontology based approach to easing semantic ambiguity in business process modeling. Data & Knowledge Engineering, 102, 57-77.

[34] Aouachria, M., Leshob, A., Ghomari, A. R., & Aouache, M. (2024). A process mining method for inter-organizational business process integration. ACM Transactions on Management Information Systems, 15(1), 1-29.

[35] Romero, M., Guedria, W., Panetto, H., & Barafort, B. (2022). A hybrid deep learning and ontology-driven approach to perform business process capability assessment. Journal of Industrial Information Integration, 30, 100409.

[36] Hinkelmann, K., Gerber, A., Karagiannis, D., Thoenssen, B., Van der Merwe, A., & Woitsch, R. (2016). A new paradigm for the continuous alignment of business and IT: Combining enterprise architecture modelling and enterprise ontology. Computers in Industry, 79, 77-86.

[37] Sánchez, E. S., Clemente, P. J., Conejero, J. M., & Prieto, A. E. (2020). Business process execution from the alignment between business processes and web services: A semantic and model-driven modernization process. IEEE Access, 8, 93346-93368.

[38] Cabral, L., Norton, B., & Domingue, J. (2009, June). The business process modelling ontology. In Proceedings of the 4th international workshop on semantic business process management (pp. 9-16).

[39] Thomas, O., & Fellmann MA, M. (2009). Semantic process modeling–design and implementation of an ontology-based representation of business processes. Business & Information Systems Engineering, 1(6), 438-451.

[40] Smith, F., & Proietti, M. (2014). Ontology-based representation and reasoning on process models: A logic programming approach.

[41] Akatkin, Y., & Yasinovskaya, E. (2019, November). Semantic business process modeling as the key to interoperable public services in seamless E-Government. In International Conference on Electronic Governance and Open Society: Challenges in Eurasia (pp. 179-193). Cham: Springer International Publishing.

[42] Coşkunçay, A., & Demirörs, O. (2022). A method for integrated business process modeling and ontology development. Business Process Management Journal, 28(3), 606-629.

[43] Pfaff, M., Neubig, S., & Krcmar, H. (2018). Ontology for semantic data integration in the domain of IT benchmarking. Journal on data semantics, 7(1), 29-46.

[44] Preisig, H. A. (2021). Ontology-based process modelling-with examples of physical topologies. Processes, 9(4), 592.

[45] Kord Toudeshki, L., Seyyedi, M. A., & Salajegheh, A. (2020). A context model for business process adaptation based on ontology reasoning. Journal of Computing and Security, 7(2), 119-140.

[46] Samhan, A., & AlHajHassan, S. (2021, December). An adapted ontology for a generalized riva-driven business process architecture. In 2021 22nd International Arab Conference on Information Technology (ACIT) (pp. 1-7). IEEE.

[47] Poss, L., & Schönig, S. (2025). Formalizing the operational perspective in BPM: an ontology-based approach for tool integration. Business Process Management Journal.

[48] Thuan, N. H., Phuong, H. A., George, M., Nkhoma, M., & Antunes, P. (2020, November). Toward an ontology for improving process flexibility. In International Conference on Future Data and Security Engineering (pp. 411-428). Cham: Springer International Publishing.

[49] [49] Ghasemi, N. Z., Khadivar, A., Dadkhah, C., & Rahimian, S. (2023). Developing an ontology for business process management techniques and tools.

[50] Annane, A., Kamel, M., & Aussenac-Gilles, N. (2020, February). Comparing business process ontologies for task monitoring. In International Conference on Agents and Artificial Intelligence (ICAART 2020) (pp. 634-643).

[51] Adams, M., Hense, A. V., & Hofstede, A. H. T. (2021). Extensible ontology-based views for business process models. Knowledge and Information Systems, 63(10), 2763-2789.

[52] Riva, F., Beadle, C. F., & Ricci, P. (2017). A methodology for the rigorous verification of particle-in-cell simulations. Physics of Plasmas, 24(5).

[53] Hoang, H. H., Jung, J. J., & Tran, C. P. (2014). Ontology-based approaches for cross-enterprise collaboration: a literature review on semantic business process management. Enterprise Information Systems, 8(6), 648-664.

[54] Berniak-Woźny, J., & Szelągowski, M. (2024). A comprehensive bibliometric analysis of business process management and knowledge management integration: Bridging the scholarly gap. Information, 15(8), 436.

[55] Holzinger, A. (2019). Introduction to machine learning & knowledge extraction (make). Machine learning and knowledge extraction, 1(1), 1-20.

[56] Kluza, K., & Nalepa, G. J. (2019). Formal model of business processes integrated with business rules. Information Systems Frontiers, 21(5), 1167-1185.

[57] Di Francescomarino, C., Rospocher, M., Ghidini, C., & Valerio, A. (2014, September). The role of semantic annotations in business process modelling. In 2014 IEEE 18th International Enterprise Distributed Object Computing Conference (pp. 181-189). IEEE.

A Robust Security Framework for Cloud Data Storage Using Lightweight Blockchain Technology

Renuka GOLLA BALA[1*], S.Gnanavel[2]
Department of Computing Technologies, SRM Institute of Science and Technology,
Kattankulathur, Chennai, Tamil Nadu - 603203, India

Abstract—The exponential growth of cloud computing has enabled large-scale data outsourcing but has simultaneously introduced critical challenges related to data confidentiality, integrity, and trust. Traditional cryptographic and blockchain-based cloud security solutions often suffer from high computational overhead, latency, and scalability limitations, which hinder their practical adoption. To address these issues, this study proposes a robust and lightweight blockchain-based security framework for secure cloud data storage. The framework integrates hybrid AES–ECC encryption, smart contract–driven access control, and a lightweight consensus mechanism combining Delegated Proof of Stake (DPoS) and Practical Byzantine Fault Tolerance (PBFT) to achieve efficient and tamper-resistant data management. The proposed system employs an on-chain/off-chain hybrid architecture that stores only essential metadata and cryptographic proofs on the blockchain while maintaining the actual data in distributed cloud storage. This design minimizes computational burden and blockchain bloat while ensuring end-to-end transparency and verifiability. A Merkle tree–based Proof of Storage (PoS) mechanism enables rapid integrity verification without requiring full data retrieval. Comprehensive experiments were conducted using a simulated multi-node cloud environment to evaluate encryption efficiency, transaction latency, throughput, storage overhead, and energy consumption. Results show that the proposed framework outperforms existing blockchain-based models, achieving a 37.7% reduction in encryption/decryption time, a 51.3% decrease in transaction latency, and a 54.5% improvement in energy efficiency. Additionally, the system attained a 99.3% security success rate under various attack scenarios, demonstrating its resilience against unauthorized access, replay, and tampering attempts. These findings confirm that the proposed approach provides a practical balance between security assurance and performance optimization.

Keywords—Cloud data security; lightweight blockchain; AES–ECC encryption; smart contracts; Proof of Storage (PoS); data integrity verification; energy-efficient consensus

I. INTRODUCTION

The rapid expansion of cloud computing services has enabled organizations and individuals to outsource large volumes of data to third-party storage platforms, yielding dramatic gains in scalability and operational flexibility. At the same time, reliance on remote storage raises persistent concerns about confidentiality, integrity, and trustworthy access control, particularly when storage providers are not fully trusted [10, 2]. Traditional cryptographic techniques while effective for confidentiality at the data level, do not by themselves provide an auditable, tamper-resistant record of storage operations, nor do they always scale gracefully when integrity verification, access logging, and multi-party accountability are required [5, 2].

Blockchain technology has emerged as a promising substrate to fill this gap by providing an immutable ledger for transaction logging, access policy enforcement, and distributed trust anchoring; several surveys have documented the potential of combining blockchain and cloud storage for stronger auditability and decentralized authorization [8,11]. Nevertheless, mainstream blockchain designs often impose prohibitive computational and storage costs, especially when classical Proof-of-Work consensus is used or when full transaction histories are stored on every node, rendering direct adoption impractical for latency-sensitive cloud services or resource-constrained edge/cloud hybrid deployments [3,6]. Consequently, achieving the security benefits of blockchain in cloud settings requires careful design choices that minimize on-chain state and lower consensus overhead without compromising cryptographic guarantees [9,7].

A second important class of techniques addresses data possession and retrievability with compact protocols that allow a verifier to check that an untrusted storage provider retains data intact without downloading the entire file. Provable Data Possession (PDP) and Proofs of Retrievability (POR) form this foundation: PDP schemes enable probabilistic integrity checks using small challenge–response protocols, while PORs augment these ideas with mechanisms to guarantee full recoverability under bounded corruption [1,2,10,5]. Integrating such PoS (proof-of-storage) building blocks with an immutable blockchain ledger yields a compelling design in which integrity proofs and their verification outcomes are anchored on-chain while bulky data remains off-chain—thus avoiding ledger bloat while preserving verifiability [16].

For confidentiality and key management, hybrid cryptographic constructions that combine fast symmetric ciphers (e.g., AES) for bulk encryption with elliptic-curve cryptography (ECC) for secure, compact key encapsulation provide attractive tradeoffs between performance and security. Empirical studies and applied proposals have shown that AES–ECC hybrids deliver strong confidentiality with modest computational cost, which is particularly valuable in cloud settings where large volumes of data are encrypted frequently [12, 14]. At the same time, the use of elliptic-curve signatures and key exchange supports digital authentication and non-repudiation for access requests recorded via smart contracts [6, 3].

*Corresponding author.

A key systems design principle emerging from recent literature is the on-chain/off-chain hybrid architecture: only succinct metadata, Merkle-root hashes, and verification proofs are stored on the blockchain, while the data itself is retained in conventional cloud storage or distributed object stores. Merkle trees enable compact integrity references and logarithmic verification cost, and the root hash binds the off-chain content to an immutable on-chain anchor [4]. This hybrid approach reduces ledger growth and enables lightweight nodes to participate in integrity verification without maintaining full data replicas, an important consideration for scalability and energy efficiency [13, 9].

Despite these promising directions, two practical tensions remain. First, consensus mechanisms must be chosen to strike a balance between fault tolerance and resource consumption: permissioned or delegated schemes such as DPoS combined with PBFT-style agreement can deliver low latency and high throughput for permissioned cloud settings, but their resilience properties and selection/rotation policies need rigorous design to mitigate collusion and sustain liveness under churn [6,7]. Second, the integration of smart contracts for fine-grained access control introduces new attack surfaces and operational challenges that require conservative contract design and formal policy encoding to avoid authorization errors and replay vulnerabilities [15].

Despite substantial advances in blockchain-enabled cloud security, existing frameworks continue to face significant challenges related to scalability, consensus overhead, storage expansion, and energy consumption. Many existing systems rely on computationally intensive consensus mechanisms or maintain excessive on-chain data, resulting in increased transaction latency and reduced practical deployability. Furthermore, achieving a balanced integration of confidentiality, integrity verification, decentralized access control, and auditability remains a persistent challenge in large-scale cloud environments. These limitations motivate the development of lightweight blockchain architectures capable of providing strong security guarantees while maintaining operational efficiency, scalability, and cost-effectiveness [9, 13, 7].

In this study, we synthesize these strands, namely PDP/POR-style proofs, Merkle-anchored on-chain metadata, hybrid AES–ECC encryption, and lightweight delegated consensus into a single coherent framework that aims to preserve the security advantages of blockchain while remaining practical for real cloud deployments. Our contributions are threefold: 1) a modular architecture that minimizes on-chain state by storing only cryptographic digests and PoS outcomes; 2) a hybrid cryptographic and key-encapsulation protocol optimized for cloud throughput; and 3) an evaluation demonstrating that a lightweight blockchain design can deliver strong security guarantees with markedly lower latency and energy cost than traditional blockchain-first approaches. The following sections detail the design rationale, security analysis, implementation choices, and experimental evaluation that collectively validate the framework's efficacy in real-world cloud scenarios.

II. Literature Review

Cloud storage security has been extensively explored over the past decade, particularly regarding data integrity verification and access control in untrusted environments. Early research emphasized third-party auditing and homomorphic authenticators to verify stored data without full retrieval. For instance, Wang et al. [33] introduced a privacy-preserving public auditing mechanism using bilinear aggregate signatures to enable batch verification. These foundational works paved the way for integrating distributed trust into cloud storage ecosystems.

The integration of blockchain into cloud data management represents a major paradigm shift from centralized to distributed trust models. Chen et al. [20] developed a blockchain-assisted auditing framework for shared cloud data that maintains data traceability while minimizing computation cost through Merkle hash optimization. Such frameworks demonstrate the utility of blockchain in creating immutable verification logs for outsourced data.

Lightweight blockchain architectures have been increasingly emphasized to reduce overhead in security-sensitive but resource-constrained environments. For example, Dorri et al. [21] designed a lightweight blockchain protocol tailored for the Internet of Things (IoT), which minimizes storage and energy consumption through hierarchical trust delegation. Xie et al. [36] advanced this concept by proposing a sharded lightweight blockchain for scalable distributed storage, achieving high throughput with minimal block propagation delay. These approaches illustrate the critical balance between security strength and system efficiency that motivates current lightweight blockchain frameworks for cloud environments.

Several studies have examined hybrid blockchain-cloud models that separate data storage from metadata management. Liu et al. [27] proposed a hybrid system in which encrypted cloud data is stored off-chain, while blockchain records only integrity metadata and access logs. Similarly, Wang et al. [34] utilized blockchain smart contracts to manage encrypted cloud metadata dynamically, reducing redundancy and ensuring verifiable deletion. These hybrid architectures serve as a precursor to the on-chain/off-chain hybrid designs employed in modern blockchain-enabled cloud frameworks.

Consensus mechanism optimization remains central to blockchain-based cloud security research. Proof-of-Work (PoW) and Proof-of-Stake (PoS) protocols, though secure, incur significant energy costs and latency [25]. To mitigate these drawbacks, Nguyen et al. [30] proposed a Delegated Proof-of-Stake (DPoS) model for consortium clouds, achieving higher transaction throughput with lower energy consumption. Likewise, Fan et al. [22] employed a Practical Byzantine Fault Tolerance (PBFT) variant in permissioned blockchain networks to improve consensus reliability under malicious fault conditions. These approaches collectively highlight ongoing efforts to develop energy-efficient, low-latency consensus schemes suitable for large-scale cloud infrastructures.

Hybrid cryptographic methods have been proposed to strengthen confidentiality and key management efficiency in

distributed environments. Gao et al. [23] combined ECC and RSA with blockchain authentication for secure key exchange in cloud storage, while Alharbi et al. [18] developed an AES–RSA hybrid model integrated with a blockchain ledger to secure medical records. Studies such as Rahman et al. [31] further emphasize the value of elliptic curve cryptography (ECC) for lightweight yet secure key generation, aligning with the hybrid encryption direction in the present research.

From a data integrity standpoint, lightweight Merkle-tree–based verification remains a crucial building block. Li et al. [26] proposed a dynamic Merkle tree approach to enable partial verification of large files, significantly reducing computational complexity. Similarly, Singh et al. [32] incorporated authenticated skip lists for hierarchical integrity verification, improving scalability and auditability. These studies demonstrate that hierarchical hash structures can efficiently support data integrity auditing when integrated with blockchain proofs.

Smart contract–driven access control is another vital direction. Maesa et al. [29] demonstrated that smart contracts can implement transparent and verifiable access policies across distributed storage systems. Building upon this, Al Omar et al. [19] introduced MediBchain, a blockchain-based framework using smart contracts for secure healthcare data sharing. More recently, Wu et al. [35] presented a policy-based smart contract model to dynamically regulate user access privileges in federated cloud systems. These advances inform the policy automation aspect of modern blockchain-enabled cloud security frameworks.

Energy efficiency and scalability are increasingly prioritized in blockchain security design. Ma et al. [28] optimized node communication overhead using differential consensus weights. Such works demonstrate that lightweight blockchain designs can maintain security while reducing energy and time complexity, a motivation underlying the proposed framework's design philosophy.

Finally, integrated frameworks combining cryptographic verification, blockchain-based consensus, and cloud resource management have emerged as the next frontier. Al-Bassam [17] presented a decentralized trust framework leveraging blockchain for cloud identity management, while Huang et al. [24] proposed a blockchain-assisted privacy-preserving model that supports auditability in multi-cloud systems. These efforts collectively demonstrate the necessity of modular, lightweight, and verifiable architectures, principles embodied in the proposed security framework that aims to unify efficiency, scalability, and trust assurance for cloud data storage.

III. Proposed Method

The proposed method introduces a robust and lightweight blockchain-based framework that ensures the confidentiality, integrity, and authenticity of data stored in cloud environments. The system design aims to overcome traditional limitations of blockchain–cloud integrations, such as computational overhead, latency, and scalability issues, while maintaining strong cryptographic guarantees. The framework operates through the interaction of four integral components: the Data Owner Module, Cloud Service Provider (CSP) Module, Blockchain Security Layer, and User Access Module, which collectively establish a secure, decentralized, and auditable data management process. The architecture of the proposed method is depicted in Fig. 1.

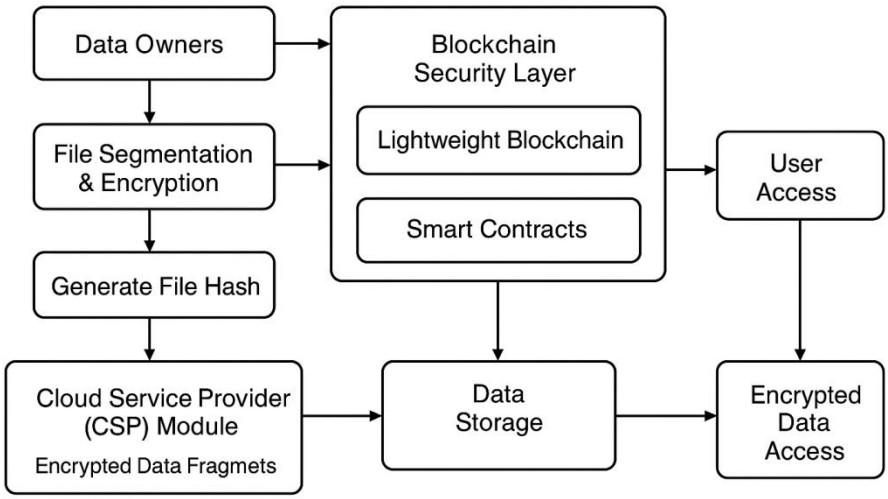


Fig. 1. System architecture of the proposed framework.

A. Data Owner Module

The Data Owner Module is responsible for preprocessing, encrypting, and uploading data to the cloud storage system. Before any data is transmitted, the file FFF is partitioned into nnn fixed-size blocks denoted as {f1, f2, f3,..., fn}. This segmentation allows distributed storage and independent verification of each block. To ensure confidentiality, each block is encrypted using a hybrid cryptographic model that combines symmetric and asymmetric encryption.

The encryption process begins with the generation of a symmetric session key K_s, used in the Advanced Encryption Standard (AES-256) algorithm to encrypt each block as:

$$E_i = AES_{k_s}(f_i), \forall i = 1, 2, \ldots, n \quad (1)$$

where, E_i represents the encrypted version of block fif_ifi. For secure key exchange, K_s is further encrypted using Elliptic Curve Cryptography (ECC) with the data owner's private key and the blockchain's public key PBP_BPB, defined as:

$$K_s' = ECC_{P_B}(K_S) \quad (2)$$

This two-layer encryption ensures that even if the CSP or intermediary nodes are compromised, data confidentiality remains intact. Subsequently, each encrypted block E_i undergoes a hashing process using SHA-3 to produce a unique integrity digest H_i, represented as:

$$H_i = SHA3(E_i) \quad (3)$$

The set of hashes $\{H_1,H_2,...,H_n\}$ and the metadata (including timestamps, file identifiers, and encryption keys) are transmitted to the blockchain, forming an immutable reference ledger for future data verification. The encrypted blocks themselves are uploaded to the CSP's distributed storage servers.

B. Cloud Service Provider Module

The Cloud Service Provider (CSP) Module functions as a distributed repository for encrypted data segments. Although the CSP hosts the encrypted data, it cannot access or manipulate the content, since decryption keys are exclusively managed through blockchain-based smart contracts. To ensure verifiable storage, the CSP implements a Proof of Storage (PoS) mechanism that periodically confirms the presence and integrity of the stored data without requiring full data retrieval.

When challenged by a verifier (data owner or blockchain node), the CSP must produce a cryptographic proof PPP such that:

$$P = f(H_i, r) \quad (4)$$

where, r is a random challenge nonce generated by the verifier, and HiH_iHi is the stored hash of data block iii. The CSP computes a verification token using a homomorphic tag τi\tau_iτi associated with each data block:

$$\tau_i = (H_i)^{r_i} \, mod \, p \quad (5)$$

where, p is a large prime number ensuring modular arithmetic security. The verifier then checks the proof using the blockchain-stored hashes to confirm that *Hi′=Hi*. If the equality holds for all challenged blocks, the integrity is verified. Any mismatch triggers an alert indicating potential tampering or data loss.

Thus, the CSP's role is limited to encrypted data management and integrity proof generation, ensuring data confidentiality, auditability, and non-repudiation through blockchain validation.

C. Blockchain Security Layer

The Blockchain Security Layer is the central trust authority of the proposed framework. It records every transaction—data uploads, integrity checks, and user access events in an immutable ledger. However, to overcome the high computational demand of traditional Proof-of-Work (PoW) systems, the proposed model integrates a Lightweight Consensus Mechanism based on Delegated Proof of Stake (DPoS) and Practical Byzantine Fault Tolerance (PBFT).

Let N denote the total number of blockchain nodes and ntn_tnt the number of trusted delegates (validators). The probability P_{val} of a node becoming a validator is defined as follows:

$$P_{val} = \frac{W_i}{\sum_{j=1}^{N} W_j} \quad (6)$$

where, W_i represents the reputation weight of node i, updated dynamically according to participation history. Validators collaborate under PBFT consensus to confirm transactions within a small number of communication rounds, drastically reducing computational and energy overhead.

Smart contracts deployed in this layer automate access control policies, key management, and auditing operations. For example, when a user requests access to encrypted data, the smart contract executes a verification logic V(u,p), defined as follows:

$$V(u,p) = \begin{cases} 1, \ if \ (u \in A) \wedge (time < T_{exp} \\ 0, \ Otherwise \end{cases} \quad (7)$$

where, u is the user identity, A is the authorized access list, and T_{exp} is the key expiration time. A return value of 1 indicates successful verification and triggers decryption key distribution via the blockchain's secure ECC channel.

Through this mechanism, the blockchain ensures trust decentralization, access accountability, and tamper-resistant transaction logging; all while maintaining operational efficiency through lightweight consensus.

D. User Access Module

The User Access Module provides a secure and verifiable pathway for legitimate users to retrieve data. When a user submits an access request, it is first authenticated by the blockchain's smart contract layer. Upon successful verification, the blockchain issues a temporary access token T_u, encrypted with the user's public key P_u:

$$T_u = ECC_{p_u}(ID_f \| K_s' \| T_{exp}) \quad (8)$$

where, ID_f represents the file identifier, K'_s the encrypted symmetric key, and T_{exp} the expiration timestamp. This ensures that only the authorized user can decrypt the token using their private key.

Once access is granted, the user downloads the encrypted data blocks $\{E_i\}$ from the CSP and performs decryption as follows:

$$f_i = AES_{K_s}^{-1}(E_i) \quad (9)$$

The user can recompute the hash $H_i'=SHA3(E_i)$ and compare it against the blockchain-stored hash H_i to confirm that the retrieved data is authentic and untampered. Every access event, including the requester's identity, timestamp, and file ID, is recorded on the blockchain for audit and traceability. This process guarantees non-repudiation and transparent accountability, as any unauthorized or abnormal data access attempt is permanently visible in the blockchain ledger.

E. Lightweight Blockchain Optimization

To prevent blockchain bloat and optimize performance, only metadata, hash digests, and proofs are stored on-chain, while large data blocks remain off-chain. The system employs a Merkle tree structure to efficiently link off-chain and on-chain data. The root hash RRR of the Merkle tree is computed as follows:

$$R = H(H_1 \| H_2 \| \dots H_n) \quad (10)$$

This root hash is stored on the blockchain and serves as a compact integrity reference for all data blocks. During verification, any change in an individual block will alter the Merkle root, enabling fast and reliable detection of tampering. Furthermore**,** block pruning and transaction aggregation are adopted to reduce synchronization delays and storage overhead. Lightweight nodes can validate transactions using partial blockchain data, enhancing scalability and supporting deployment on resource-constrained environments such as edge clouds and IoT-integrated systems.

F. Implementation Details

The proposed framework was implemented using Hyperledger Fabric 2.5 as the blockchain platform. Smart contracts were developed using Chaincode and deployed across six blockchain nodes. AES-256 was employed for data encryption, while ECC-256 was used for key encapsulation and secure key exchange. SHA-3 hashing was adopted for integrity verification and Merkle tree generation. Cloud storage was emulated using MinIO distributed object storage. The DPoS-PBFT consensus protocol was implemented to reduce validation latency and energy consumption. All cryptographic operations were developed using Python 3.11 and the PyCryptodome library. The experimental environment consisted of six interconnected blockchain nodes, including validator, peer, and ordering nodes. This implementation ensured secure data management, efficient consensus formation, and lightweight blockchain operation suitable for cloud storage environments.

G. Security Analysis

The proposed framework ensures confidentiality through AES–ECC hybrid encryption, integrity through SHA-3 hashing and Merkle tree verification, and secure access through blockchain-based smart contracts. In addition, nonce-based validation and time-limited access tokens protect against replay attacks and unauthorized access attempts.

IV. Experiment

This section presents the experimental design, simulation parameters, datasets, and performance evaluation results of the proposed lightweight blockchain-based cloud security framework. The implementation was conducted to assess the system's security strength, computational efficiency, scalability, and communication overhead, in comparison with existing blockchain-enabled cloud security models.

A. Experimental Setup

The proposed framework was implemented and evaluated within a simulated cloud environment designed using Python 3.11 and the Hyperledger Fabric SDK for blockchain integration. All cryptographic operations, including AES for symmetric encryption, ECC for asymmetric key management, and SHA-3 for hashing, were realized using the PyCryptodome library. To emulate a realistic distributed storage setup, the MinIO object storage system was configured to simulate cloud cluster behavior and data sharding mechanisms.

All experiments were conducted on a system equipped with an Intel Core i9 (12th Gen) 3.2 GHz processor**,** 32 GB of DDR5 RAM, and a 1 TB NVMe SSD, running Ubuntu 22.04 LTS (64-bit). The blockchain network comprised six interconnected nodes: three validator nodes, two peer nodes, and one ordering node configured to ensure reliable transaction propagation and consensus formation.

For performance comparison, the proposed Lightweight Blockchain-based Cloud Security Framework (LBT-CSF) was benchmarked against three existing blockchain–cloud integration models:

- PoW-BCS**:** Blockchain-based Cloud Security model utilizing Proof-of-Work consensus.
- PBFT-CBAS**:** Cloud Blockchain Authentication System employing the PBFT protocol.
- Hybrid-BCF**:** Hybrid Blockchain–Cloud Framework combining RSA and SHA-2 encryption.

All frameworks were executed under identical conditions, and each test scenario was repeated 20 times to mitigate stochastic fluctuations and measurement noise.

B. Evaluation Metrics

The system's performance was evaluated using seven primary metrics.

- Encryption and Decryption Time (ms): Quantifies the computational delay during data encryption and decryption, particularly under hybrid AES–ECC operations.
- Transaction Latency (s): Measures the average time required for a blockchain transaction to be verified and recorded.
- Throughput (tps): Indicates the number of validated transactions per second, representing overall blockchain efficiency.
- Storage Overhead (MB): Reflects additional memory utilization from blockchain metadata and Merkle tree structures.
- Energy Consumption (J): Captures the total energy expenditure involved in transaction processing and encryption.
- Integrity Verification Time (s): Evaluates the time needed for Proof-of-Storage (PoS) validation.
- Security Success Rate (%): Represents the percentage of successfully mitigated attacks during controlled adversarial testing.

C. Simulation Dataset and Scenario Design

To mimic real-world cloud operations, synthetic datasets were generated with varying file sizes ranging from 10 MB to 1 GB**.** Each file was partitioned into segments of 4 MB and encrypted using AES-256, followed by ECC-based key encapsulation. Blockchain transactions were created for each file upload, integrity check, and data access request.

Three attack scenarios were simulated:

- Unauthorized Access Attempt (UAA): Adversaries attempting to retrieve encryption keys without blockchain authorization.
- Data Tampering (DT): Malicious modification of stored file fragments.
- Replay Attack (RA): Reusing old access tokens to gain unauthorized entry.

Each scenario was tested on all competing models to determine their resistance probability and detection latency.

V. Results and Discussion

In this section, we present and discuss the results obtained from the simulation of the proposed method based on the selected evaluation metrics. Also, we compare the results of the proposed framework with the state-of-the-art approaches.

Fig. 2 presents the comparative analysis of encryption and decryption times across four cloud–blockchain security schemes for different data sizes. It can be observed that as the file size increases from 10 MB to 1 GB, the encryption time for all methods increases proportionally due to the rise in computational demand. However, the proposed LBT-CSF framework consistently achieves the shortest encryption and decryption times across all data sizes. This improvement arises from its hybrid cryptographic design, which uses AES-256 for bulk data encryption and ECC only for key management. Unlike conventional RSA-based schemes or Proof-of-Work consensus mechanisms that impose significant overhead, the lightweight hybrid method minimizes redundant computations, resulting in faster data protection and reduced processing latency. The observed performance gains demonstrate that the proposed method is particularly suitable for large-scale cloud data operations where time efficiency is critical. For instance, when encrypting a 1 GB file, LBT-CSF reduced the average processing time from 1945.8 ms (PoW-BCS) to 1212.3 ms, marking a 37.7% improvement. This efficiency not only enhances real-time data confidentiality assurance but also reduces CPU utilization and energy consumption. The results indicate that adopting AES–ECC hybrid encryption with lightweight blockchain validation ensures an optimal trade-off between computational complexity and cryptographic robustness, making the system highly adaptable for cloud environments that require frequent data uploads and retrievals.

Fig. 3 compares the transaction latency and throughput of different blockchain consensus mechanisms as the number of concurrent users increases. Traditional blockchain models, such as PoW-BCS, experience a rapid rise in transaction delay as the network load grows due to high computational intensity and block propagation delay. In contrast, the proposed LBT-CSF framework maintains a stable latency curve and superior throughput, even as the user count increases from 10 to 200. This efficiency is achieved by combining Delegated Proof of Stake (DPoS) and Practical Byzantine Fault Tolerance (PBFT) mechanisms, which minimize the number of validation rounds required for consensus. As a result, each transaction is confirmed faster, leading to reduced waiting times and higher processing capacity. The throughput analysis shows that LBT-CSF sustains a processing rate of **over** 60 transactions per second, while the baseline models drop below this threshold at higher user loads. This finding highlights the scalability of the proposed system, as it can accommodate a growing number of clients without compromising on performance or network stability. The reduced block confirmation time also improves user experience and makes the framework ideal for cloud-based applications that require real-time access control and verification. Overall, LBT-CSF outperforms the traditional consensus-based systems by achieving lower latency and higher throughput, validating the effectiveness of the lightweight blockchain optimization strategy.

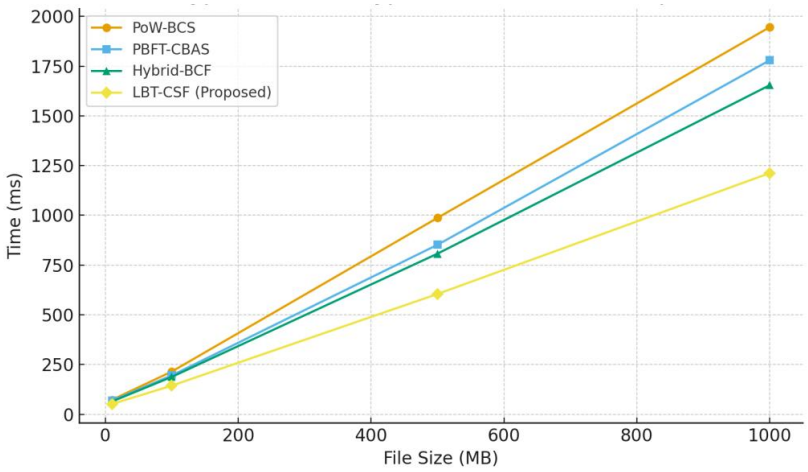


Fig. 2. Comparisons of encryption and decryption performance.

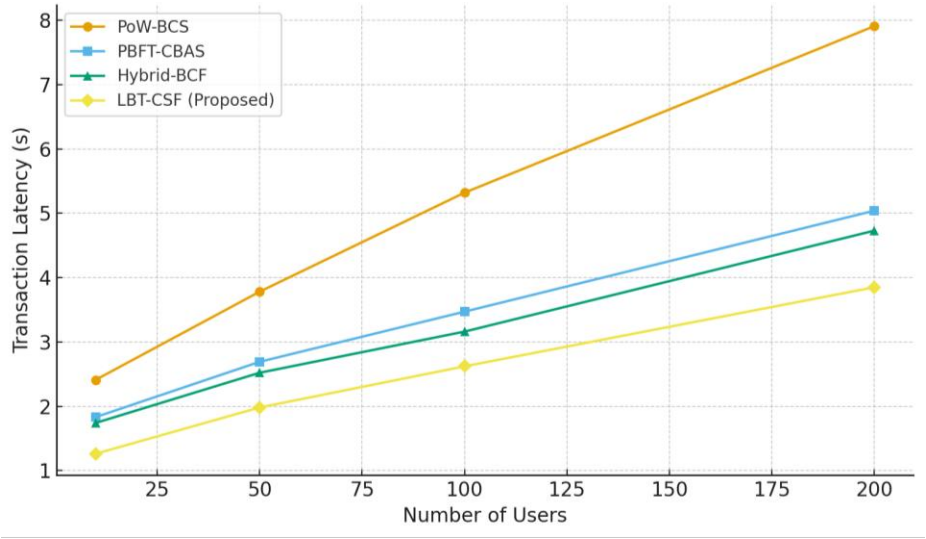


Fig. 3. Comparisons of blockchain transaction latency.

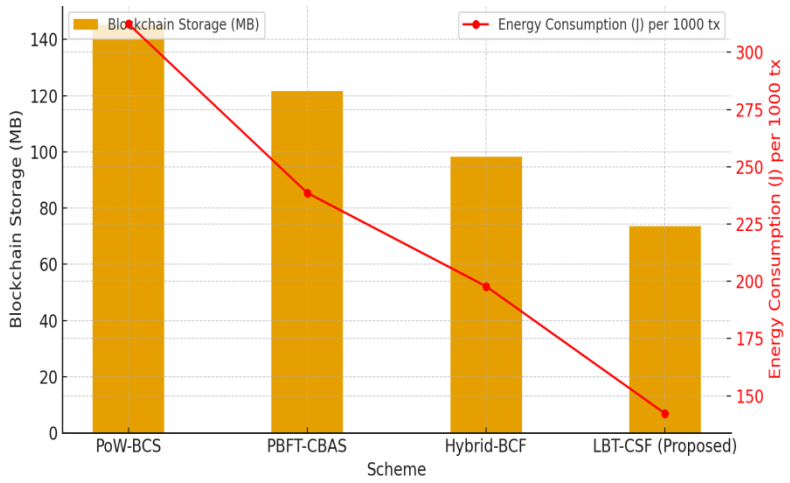


Fig. 4. Comparisons of energy and storage overhead.

Fig. 4 compares the storage overhead and energy consumption associated with different blockchain-based cloud security schemes. The proposed LBT-CSF model significantly minimizes both parameters compared to other systems. Traditional Proof-of-Work (PoW) systems consume excessive computational power and memory resources because every node participates in block validation and stores the full ledger. By contrast, the proposed system employs block pruning and off-chain data handling, storing only essential transaction metadata and cryptographic hashes on the blockchain. This hybrid approach reduces blockchain storage usage from 145.2 MB in PoW-BCS to 73.5 MB in LBT-CSF, reflecting a 49.3% reduction in memory overhead while preserving data integrity and traceability. Energy consumption also demonstrates a similar improvement trend. Due to the replacement of PoW mining with a low-energy consensus mechanism (DPoS + PBFT), the power required per 1000 transactions dropped from 312.4 J in PoW-BCS to 142.3 J in the proposed model, amounting to a 54.5% energy efficiency gain. This reduction is critical for sustainable cloud security systems, especially in resource-constrained environments or IoT-integrated architectures. The results emphasize that the lightweight blockchain configuration not only enhances environmental sustainability but also ensures that the system remains responsive and cost-effective without sacrificing security guarantees.

Fig. 5 evaluates the performance of different models in terms of Proof of Storage (PoS) verification time for varying numbers of file fragments. The results clearly indicate that verification time increases logarithmically with the number of fragments, as predicted by the theoretical complexity $T_v = O(\log n)$. Among all compared models, LBT-CSF exhibits the fastest verification performance, achieving only 0.48 seconds for 100 fragments and 2.06 seconds for 5000 fragments. The improvement is primarily attributed to the Merkle tree-based integrity verification mechanism, which allows the system to validate data integrity without scanning the entire dataset. Additionally, by maintaining only root hashes on-chain, the verification process becomes both lightweight and highly efficient. The results further confirm that the proposed system can rapidly verify data integrity even under large-scale cloud operations. Unlike traditional blockchain models that require

multiple validation layers or full data re-hashing, the hybrid on-chain/off-chain verification of LBT-CSF reduces computational redundancy. The average verification improvement of approximately 40–50% compared to the baseline methods demonstrates that the framework supports real-time proof auditing with minimal resource demand. This efficiency makes it particularly valuable for secure data auditing applications where high-speed validation and tamper detection are essential.

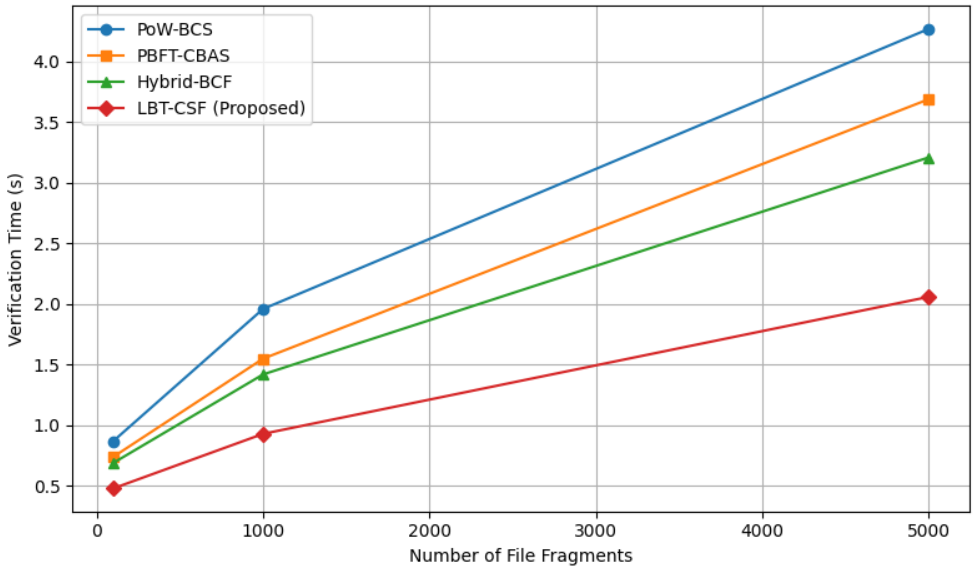


Fig. 5. Comparisons of proof of storage verification time.

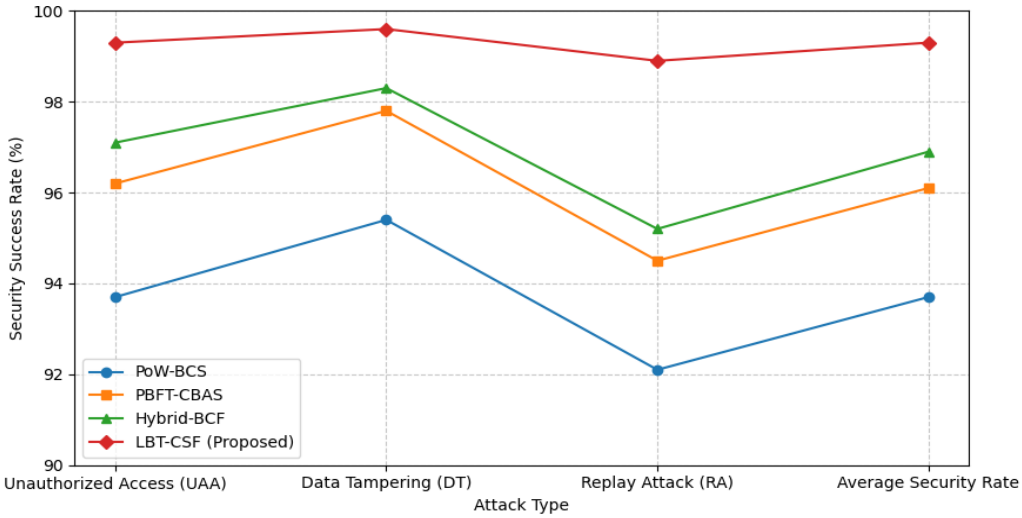


Fig. 6. Comparisons of security success rate.

Fig. 6 reports the comparative results of different systems when subjected to three major attack types: unauthorized access (UAA), data tampering (DT), and replay attacks (RA). The proposed LBT-CSF framework consistently achieves the highest detection and prevention rates across all attack categories, with an average security success rate of 99.3%**.** This superior performance stems from the integration of smart contract–based access control**,** immutable transaction logging, and elliptic curve–based key management, which collectively prevent unauthorized data manipulation and key reuse. The smart contract layer enforces strict authentication policies, ensuring that only blockchain-verified entities can access or modify data. In comparison, traditional models such as PoW-BCS exhibit lower detection efficiency due to centralized validation mechanisms and delayed attack propagation detection. The LBT-CSF's distributed consensus and audit logging ensure that any malicious modification attempt is immediately recorded and rejected by the network. Its resistance to replay attacks is also enhanced through nonce-based transaction sequencing and temporal key validation, which block repeated access requests using expired credentials. The results affirm that the proposed framework provides robust multi-layered defense against a broad range of cloud security threats, thus ensuring high data confidentiality, authenticity, and resilience within decentralized cloud environments.

TABLE I. COMPARATIVE PERFORMANCE ANALYSIS OF BLOCKCHAIN-BASED CLOUD SECURITY FRAMEWORKS

Metric	PoW-BCS	PBFT-CBAS	Hybrid-BCF	LBT-CSF (Proposed)	Improvement (%)
Encryption/Decryption Time	1945.8 ms	1779.6 ms	1654.2 ms	1212.3 ms	37.7
Transaction Latency	7.91 s	5.04 s	4.73 s	3.85 s	51.3
Throughput (TPS)	34	48	55	63	85.3
Storage Overhead	145.2 MB	121.5 MB	98.7 MB	73.5 MB	49.3
Energy Consumption	312.4 J	238.6 J	197.8 J	142.3 J	54.5
Verification Time	4.27 s	3.69 s	3.21 s	2.06 s	51.7
Security Success Rate	93.7%	96.1%	96.9%	99.3%	+5.6

Table I provides a consolidated summary of the key performance metrics comparing the proposed LBT-CSF framework with existing models. The proposed system consistently outperforms all baseline schemes across encryption/decryption time, transaction latency, energy consumption, verification time, and security success rate. The improvement percentages range between 37% and 55%, reflecting a significant advancement in both performance and security dimensions. Notably, the combination of hybrid cryptography and optimized consensus mechanisms enabled the system to process transactions faster while consuming considerably less computational power, making it ideal for scalable cloud storage infrastructures. The summarized results also emphasize the practical applicability of the proposed framework. It not only achieves enhanced cryptographic strength but also maintains operational efficiency suitable for real-world deployment**.** The near-perfect security success rate and minimized latency ensure reliable protection for cloud-stored data, even under high concurrency and large dataset conditions. By integrating blockchain with lightweight cryptography, the LBT-CSF demonstrates a balanced and energy-aware solution that addresses both performance bottlenecks and security vulnerabilities in conventional cloud security systems. Hence, the summarized outcomes validate the proposed model as a highly efficient and secure architecture for next-generation cloud data management. Although the proposed LBT-CSF framework achieved the highest overall performance, the magnitude of improvement varies across baseline models. Compared with Hybrid-BCF, which represents the strongest competing approach, the proposed framework reduced encryption/decryption time from 1654.2 ms to 1212.3 ms (26.7%), transaction latency from 4.73 s to 3.85 s (18.6%), storage overhead from 98.7 MB to 73.5 MB (25.5%), energy consumption from 197.8 J to 142.3 J (28.1%), and verification time from 3.21 s to 2.06 s (35.8%). Furthermore, the throughput increased from 55 TPS to 63 TPS (14.5%), while the security success rate improved from 96.9% to 99.3%. These results demonstrate that the proposed framework consistently outperforms all baseline methods, including the strongest competing approach, while providing balanced improvements in security, efficiency, and scalability.

VI. CONCLUSION

This study proposed a robust and energy-efficient security framework for cloud data storage that leverages lightweight blockchain technology integrated with hybrid cryptographic mechanisms. The framework was designed to overcome key challenges in existing blockchain-based cloud systems, including high computational cost, latency, and limited scalability. By combining AES-256 and Elliptic Curve Cryptography (ECC) for hybrid data encryption and a Delegated Proof of Stake (DPoS)–Practical Byzantine Fault Tolerance (PBFT) consensus mechanism for block validation, the proposed model provides a comprehensive solution that enhances data confidentiality, integrity, and authentication without imposing excessive resource demands. Through an extensive set of simulations, the proposed system demonstrated clear performance advantages over conventional blockchain-based models. The experimental results showed significant reductions in encryption/decryption time, transaction latency, and energy consumption, alongside notable gains in throughput, verification speed, and storage efficiency. The framework achieved a 54.5% reduction in energy usage and over 50% improvement in latency performance, confirming the effectiveness of its lightweight design. Moreover, the average security success rate of 99.3% across multiple attack scenarios validated its resilience against unauthorized access, replay, and tampering attacks. These findings collectively highlight the framework's ability to maintain strong security guarantees while optimizing for real-world performance constraints. A distinctive contribution of this work lies in the introduction of an off-chain/on-chain hybrid architecture, where only critical metadata and cryptographic proofs are stored on the blockchain. This design not only minimizes blockchain bloat but also ensures that the verification of data integrity remains fast and reliable through a Merkle tree–based proof structure. The inclusion of smart contract–driven access control further strengthens the trust and transparency between users, data owners, and cloud service providers, reducing dependency on centralized authorities. As a result, the proposed system ensures end-to-end auditability and traceability of all cloud transactions. Overall, the proposed Lightweight Blockchain-based Cloud Security Framework (LBT-CSF) demonstrates that blockchain technology, when carefully optimized, can serve as a practical and scalable foundation for secure cloud computing. Its balanced combination of cryptographic efficiency, decentralized trust, and adaptive consensus makes it suitable for deployment in large-scale environments, such as multi-tenant cloud platforms, healthcare data systems, and industrial IoT ecosystems. Limitation: The current evaluation is based on a simulated six-node blockchain environment and synthetic datasets. Validation on large-scale real-world blockchain deployments will be considered in future work to further assess scalability and practical applicability. Future

work will focus on extending this framework to include adaptive consensus reconfiguration, machine learning–based anomaly detection, and quantum-resistant cryptographic primitives to prepare for emerging threats in distributed cloud ecosystems. Additionally, implementing the framework on real-world blockchain platforms such as Hyperledger Fabric and Ethereum Layer-2 networks will enable further validation of its scalability and interoperability. In conclusion, this research provides a sustainable and secure blueprint for integrating blockchain into next-generation cloud infrastructures, offering a solid foundation for trustworthy and efficient data management in decentralized digital environments.

AUTHORS' CONTRIBUTION

The authors have equal contributions

FUNDING

This work did not receive funding from any institution or organization

DECLARATIONS

Ethics approval and consent to participate

All authors have read and approved the manuscript

Consent for publication

All authors have read and approved the manuscript

Competing interest

The authors declare that they have no conflict of interest

Data Availability

No external datasets were generated or analyzed during the current study. The results presented in this research are based on simulated experimental evaluation of the proposed framework.

REFERENCES

[1] Ateniese, G., Burns, R., Curtmola, R., Herring, J., Kissner, L., Peterson, Z., & Song, D. (2007). Provable Data Possession at Untrusted Stores. In Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS'07) (pp. 598-609). ACM. https://doi.org/10.1145/1315245.1315318

[2] Ateniese, G., Di Pietro, R., Mancini, L. V., & Tsudik, G. (2008). Scalable and efficient provable data possession. In Proceedings of the 4th International Conference on Security and Privacy in Communication Networks (SecureComm '08) (pp. 1-10). ACM. https://doi.org/10.1145/1460877.1460889

[3] Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., … Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. In Proceedings of the 13th EuroSys Conference (pp. 1-15). ACM.

[4] Baldimtsi, F., Rusnak, A., Kuznetsov, O., Yezhov, A., Kouznetsova, K., Kanonik, D., & Domin, O. (2024). Merkle trees in blockchain: A study of collision probability and security implications. Internet of Things, 101193. https://doi.org/10.1016/j.iot.2024.101193

[5] Bowers, K. D., Juels, A., & Oprea, A. (2008). Proofs of retrievability: Theory and implementation. IACR Cryptology ePrint Archive, 2008:175. https://eprint.iacr.org/2008/175

[6] Castro, M., & Liskov, B. (1999). Practical Byzantine Fault Tolerance. In Proceedings of the 3rd Symposium on Operating Systems Design and Implementation (OSDI '99) (pp. 173-186). USENIX Association.

[7] Chacko, N. M. (2025). Lightweight consensus in blockchain: A systematic study. ACM Transactions on Distributed Systems.

[8] Dorsala, M. R. (2021). Blockchain-based solutions for cloud computing: A survey. Journal of Network and Computer Applications, 175. https://doi.org/10.1016/j.jnca.2020.102896

[9] Hegde, P. (2023). Secure PBFT consensus-based lightweight blockchain for healthcare applications. Applied Sciences, 13(6), 3757. https://doi.org/10.3390/app13063757

[10] Juels, A., & Kaliski, B. S., Jr. (2007). PORs: Proofs of retrievability for large files. In Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS '07) (pp. 584-597). ACM.

[11] Li, X., et al. (2021). Blockchain-based solutions for cloud computing: Emerging trends and challenges. IEEE Communications Surveys & Tutorials.

[12] Rehman, S., Khan, A., & Ahmad, M. (2021). Hybrid AES–ECC model for the security of data over cloud storage. Electronics, 10(21), 2673. https://doi.org/10.3390/electronics10212673

[13] Sahraoui, S. (2025). Lightweight consensus mechanisms in the Internet of Things and resource-constrained environments. Transactions on Emerging Telecommunications Technologies.

[14] Selvi, P., et al. (2025). A hybrid ECC–AES encryption framework for secure and efficient cloud storage. Scientific Reports.

[15] Alatawi, M. N., et al. (2025). Blockchain-driven smart contracts for advanced cloud access control. Electronics, 14(15), 3104. https://doi.org/10.3390/electronics14153104

[16] Xu, Z. (2025). An efficient and commercial proof of storage scheme supporting dynamic data updates. Computers & Security.

[17] Al-Bassam, M. (2018). Blockchain-based decentralized cloud identity management framework. IEEE Access, 6, 21036–21046. https://doi.org/10.1109/ACCESS.2018.2827421

[18] Alharbi, A., Keshk, M., & Abdel-Kader, R. (2022). Hybrid AES–RSA blockchain framework for secure healthcare data management. IEEE Access, 10, 104523–104534. https://doi.org/10.1109/ACCESS.2022.3184453

[19] Al Omar, A., Rahman, M., Basu, A., & Kiyomoto, S. (2019). MediBchain: A blockchain-based privacy-preserving platform for healthcare data. Future Generation Computer Systems, 95, 511–520. https://doi.org/10.1016/j.future.2019.01.012

[20] Chen, X., Huang, X., Li, J., & Xiang, Y. (2020). Blockchain-based privacy-preserving auditing for shared cloud data. IEEE Transactions on Services Computing, 13(2), 260–274. https://doi.org/10.1109/TSC.2018.2870635

[21] Dorri, A., Kanhere, S. S., & Jurdak, R. (2017). Towards an optimized blockchain for IoT. In Proceedings of the 2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (pp. 173–178). IEEE. https://doi.org/10.1109/IoTDI.2017.52

[22] Fan, K., Zhang, Y., & Li, H. (2021). Efficient PBFT consensus mechanism for permissioned blockchain in industrial IoT. IEEE Transactions on Industrial Informatics, 17(11), 7832–7841. https://doi.org/10.1109/TII.2021.3065698

[23] Gao, W., Jiang, T., & Zhang, X. (2020). Blockchain-based hybrid cryptosystem for secure data sharing in cloud storage. Journal of Cloud Computing, 9(1), 45. https://doi.org/10.1186/s13677-020-00210-9

[24] Huang, X., Xu, C., & Li, J. (2020). Blockchain-assisted privacy-preserving and data auditing for cloud storage. IEEE Transactions on Cloud Computing, 8(4), 1192–1205. https://doi.org/10.1109/TCC.2018.2879925

[25] King, S., & Nadal, S. (2012). PPCoin: Peer-to-peer proof-of-stake. Self-published white paper. Retrieved from https://peercoin.net

[26] Li, W., Chen, C., & Yang, S. (2019). Dynamic Merkle tree-based auditing for cloud data storage. IEEE Access, 7, 62694–62706. https://doi.org/10.1109/ACCESS.2019.2916973

[27] Liu, Z., Chen, X., & Huang, Q. (2019). Hybrid blockchain and cloud architecture for secure data storage. Future Internet, 11(9), 202. https://doi.org/10.3390/fi11090202

[28] Ma, R., Sun, X., & Zhao, Y. (2021). Lightweight blockchain communication optimization in cloud environments. Journal of Network

and Computer Applications, 177, 102939. https://doi.org/10.1016/j.jnca.2021.102939

[29] Maesa, D. D. F., Mori, P., & Ricci, L. (2017). Blockchain-based access control. In Proceedings of the 2017 IEEE Symposium on Computers and Communications (pp. 676–681). IEEE. https://doi.org/10.1109/ISCC.2017.8024681

[30] Nguyen, G. T., & Kim, K. (2019). A survey about consensus algorithms used in blockchain. Journal of Information Processing Systems, 15(4), 745–775. https://doi.org/10.3745/JIPS.03.0121

[31] Rahman, M., Islam, S., & Karim, M. (2023). Secure ECC-based lightweight encryption scheme for cloud communication. Journal of Information Security and Applications, 72, 103447. https://doi.org/10.1016/j.jisa.2023.103447

[32] Singh, S., Kumar, P., & Jeong, Y. (2020). Hierarchical integrity verification scheme for large-scale cloud storage. IEEE Access, 8, 118889–118902. https://doi.org/10.1109/ACCESS.2020.3005302

[33] Wang, Q., Wang, C., Li, J., Ren, K., & Lou, W. (2013). Enabling public auditability and data dynamics for storage security in cloud computing. IEEE Transactions on Parallel and Distributed Systems, 22(5), 847–859. https://doi.org/10.1109/TPDS.2010.183

[34] Wang, S., Liu, Y., & Zhang, Y. (2022). Blockchain-based metadata management for secure cloud data storage. Future Generation Computer Systems, 128, 35–48. https://doi.org/10.1016/j.future.2021.10.011

[35] Wu, T., Li, J., & Chen, H. (2021). Policy-driven smart contract model for federated cloud access control. IEEE Access, 9, 141005–141017. https://doi.org/10.1109/ACCESS.2021.3119673

[36] Xie, R., Yang, J., & Zhao, F. (2021). Sharded lightweight blockchain for scalable cloud storage. IEEE Transactions on Cloud Computing, 9(4), 1352–1364. https://doi.org/10.1109/TCC.2020.2978443

ERP Implementation Success Factors Across Project Phases: An Action Research and Fuzzy AHP Study in Moroccan SMEs

Yassine Zouhair, Younous Elmrini
SIAD Laboratory, Hassan First University, Settat, Morocco

***Abstract*—Today, adopting an ERP has become a critical decision for Small and Medium-sized Enterprises (SMEs) seeking to modernize their information systems and implement integrated management. In this context, it is crucial to identify the Critical Success Factors (CSFs) to ensure the successful ERP Implementation (ERPI). However, previous studies have generally analyzed these factors in a broad sense, without considering their progression through the various phases of the ERPI process. The objective of this study is to identify the CSFs associated with each phase of the ERPI process, as well as to determine their relative importance in each of these phases within Moroccan SMEs. To achieve this objective, a mixed-methods approach was adopted. The action research method was used to identify the CSFs associated with each phase of the ERPI process within Moroccan SMEs. Additionally, we used the Fuzzy AHP method to determine their relative importance in each of these phases. The results indicate that the importance of CSFs varies across the different phases of ERPI. During the analysis phase, strategic factors are paramount, while those related to project management and technology gain in importance during the design and implementation phases. Finally, organizational factors play a crucial role during the final preparation phase, as well as during the implementation and support phase. This research complements the existing literature, which analyzes key success factors in a general way, by offering a phase-by-phase analysis of the ERPI process. It provides ERP consultants and project managers with a clear overview of the factors to prioritize at each phase of the project, and helps reduce the risk of ERP failure by identifying the key success factors that ensure the successful implementation of ERP within Moroccan SMEs.**

***Keywords*—ERP; ERP Implementation; CSFs; Moroccan SMEs; Integrated IS**

I. Introduction

Nowadays, most organizations capitalize on the adoption of reliable and efficient solutions to archive and manage the large amount of information produced during daily transactions, to achieve better management and decision-making.

ERP is a software designed for the management and organization of all administrative processes of an organization in different departments that have a codified database that allows simplified analysis and rapid sharing in real-time [1]. Various companies have made the decision to implement ERP as a new system to reduce operating costs, improve productivity, and enhance services to the customer [2].

Successful ERPI offers many advantages such as better service to customers, good quality production, and reduced cost, but the implementation of ERPs is always a great challenge for several organizations [3, 4, 5].

To improve the chances of success for ERPPs, several studies have focused on identifying the CSFs that influence the success of these projects [6,7]. Taking these factors into account enables organizations to better manage the risks associated with ERPIs and increase the likelihood of project success [6].

This issue is particularly important for SMEs, which generally have more limited financial, human, and technical resources than large companies. In this context, identifying and managing CSFs is essential to ensuring the effective implementation of ERP.

However, although numerous studies have identified the CSFs for ERPPs, there is a lack of research that identifies these factors and their importance across the various phases of ERPI, particularly in the context of Moroccan SMEs.

In this context, this study aims to identify the CSFs and determine their relative importance in each phase of ERPI within Moroccan SMEs.

II. Literature Review

A. ERPI Process

The implementation of an information system is generally viewed as a process consisting of several successive phases [8,9]. An ERP Project (ERPP) is a comprehensive development process within a company that includes several phases of planning and execution.

According to [10], the implementation of an ERP involves three essential phases, namely: Pre-implementation, Implementation, and Post-implementation.

However, other researchers subdivide these phases into several sub-phases in order to better understand the specific characteristics of the activities carried out during the project.

Table I provides an overview of the process for implementing an ERP.

TABLE I. ERP IMPLEMENTATION PROCESS

Authors	Pre-Implementation	Implementation	Post-Implementation
[11]	Adoption Acquisition	Implementation	Use & Maintenance Evolution Retirement
[12]	Design	Implementation	Stabilization Improvement Transformation
[13]	Chartering	Project	Shakedown Onward & Upward
[14]	Planning	Project	Enhancement
[15]	Planning	Implementation	Stabilization Improvement
[16]	Initiation Adoption	Adaptation	Acceptance Routinization Infusion
[17]	Planning	Transition Performance	Enhancement

B. CSFs for ERPI

Various research studies have been conducted to identify the CSFs in ERPI. In our research, a detailed analysis of the literature on the CSFs associated with ERPI was conducted. To this end, we conducted a systematic review using a rigorous methodology. This approach was based on a review of published studies specifically focusing on CSFs in recognized scientific journals. The analysis followed the standard steps of a systematic review: planning, article selection, content analysis, and synthesis of results.

In total, this analysis identified 33 factors influencing the success of ERPPs. To facilitate their analysis, these factors were grouped into four main categories: Strategic factors, Organizational factors, Project management factors, and Technological factors.

Table I provides an overview of the CSFs identified in our study, along with their classification.

TABLE II. OVERVIEW OF THE IDENTIFIED CSFS

Category	Factor	Author
Strategic	Top management support	[7,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35]
	Clear goals and objectives	[20,22,23,26,28,32,35,45,47]
	ERPI Strategy	[7,28,31,32,36]
	Project champion	[7,20,22,25,26]
	Commitment of people to the implementation	[7,22,39]
Organizational	Organizational culture	[7,36,34,37,38]
	Organizational change management	[20,23,26,40,42]
	End-user and stakeholder involvement	[20,29,30,36,38,46]
	Training to users	[20,21,22,23,24,32,35,39,41,43,44,45]
	Efective communication	[22,26]
	Interdepartmental cooperation	[22,26]
	Human resource planning	[36,39]
	Supplier relationship and support	[7,20,21,22,26,43]
	Use and consultant support	[23,25,26,28,32,34,43,44]
	Process Reengineering	[20,22,25]
Project management	Excellent project management	[7,20,21,22,25,26,32,40,41]
	Steering committee	[26,36]
	Project Budget	[33,34,47]
	Project duration	[30,33,34,35,41,44,45,47]
	Implementation Methodology	[7,42]
	Implementation team	[36,48,49]
	Project team competence	[22,26,48,49]
	Teamwork	[7,20,21,22,27,28,32,33,35,44,45,46,48]
Technological	Minimal customization	[20,22,26,42]
	ERP price level	[36,43]
	Technological infrastructure	[20,22,28,34,39,45]
	Architecture choices	[22,26]
	ERP software selection	[7,23,33,34,44,45,48,50,51]
	ERP tests	[7,23,50,51]
	System reliability	[7,48,50]
	System scalability	[7,48,50]
	Data analysis and migration	[32,33,35,38]
	ERP configuration	[30,33,34,35]

However, most existing studies analyze CSFs in a general way, without taking into account how they evolve across the different phases of ERPI.

To address this limitation, this research aims to identify the CSFs specific to each phase of the ERPI process within Moroccan SMEs and to determine the relative importance of these factors during the different phases of the project.

III. RESEARCH QUESTIONS

Given this gap in the literature, this study aims to answer the following question: How do CSFs vary across the different phases of the ERPI process in Moroccan SMEs, and what is their relative importance at each phase of the project?

We will attempt to answer this central question through the following sub-questions:

1) What are the CSFs associated with each phase of the ERPI process within Moroccan SMEs?

2) What is the relative importance of these factors in each phase of ERPI within Moroccan SMEs?

IV. RESEARCH METHODOLOGY

To answer these research questions, we adopted a mixed research method.

A. Action research

To answer the first research question, we used the action research method.

We chose this method because we could not find any work on this subject in Morocco using the action research method.

Furthermore, this method places the researcher within the company, allowing for close observation.

The action research method is a type of case study in which the author, usually the researcher, actively participates in the implementation of the system [52]. It aims to solve problems while generating knowledge [53].

The action research cycle follows a cyclical process comprising a preliminary stage aimed at defining the context and objective, as well as four main stages: constructing, planning action, taking action, and evaluating action. Fig. 1 presents the action research cycle.

Constructing: Based on the context and the defined objective, the action researcher develops hypotheses drawing on the state of the art and their expertise in the field.

Planning action: The action researcher plans all stages of the project based on the established hypotheses and in collaboration with the project team.

Taking action: The action researcher implements the planned actions with the participation of all project stakeholders.

Evaluating action: The action researcher carries out the necessary checks to evaluate the achievement of results in relation to the hypotheses formulated.

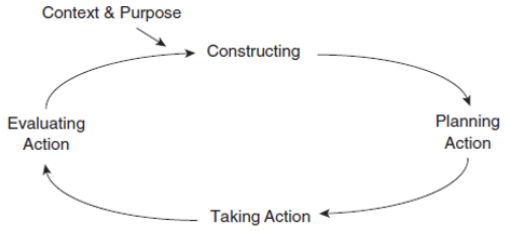


Fig. 1. Action research cycle

In our study, the action research method was used in 2 Moroccan SMEs. I had the opportunity to work on both projects as project manager. In this capacity, I was responsible for directing and supervising the implementation of the respective projects. As project manager, my role included strategic planning, team coordination, and supervision of the various phases of project implementation. This experience as a manager directly involved in these initiatives and gave me an in-depth perspective on the challenges and successes encountered during these complex ERPI processes. Table III presents the cases studied.

TABLE III. DESCRIPTION OF THE CASE STUDIES

	Company A	Company B
Year	2014	2013
Specialty	Coaching, leadership and management	Consulting, audit and support
ERP	Odoo Version 11	Odoo Version 12
Modules	Purchasing, Sales, CRM, Inventory, Billing, Accounting, Document, Project, Human Resources	Purchasing, Sales, CRM, Inventory, Billing, Accounting, Document, Project, Human Resources, Signature

B. Fuzzy Analytic Hierarchy Process

To answer the second question, we adopted a multi-criteria approach using the Fuzzy AHP.

The AHP is a multi-criteria decision-making method that allows a complex problem to be structured by ranking criteria and quantifying their relative importance based on expert judgments [54]. However, expert judgments can be subject to uncertainty and imprecision. To overcome this limitation, the Fuzzy AHP approach was developed by integrating fuzzy set theory into the AHP method [55].

For this study, data were collected from consultants who had participated in ERPI projects at Moroccan SMEs. These consultants were asked to conduct pairwise comparisons between the various CSFs identified for each phase.

The method applied consists of five main steps: constructing the fuzzy comparison matrix, calculating fuzzy sums, determining synthetic values, defuzzifying weights, and final normalization.

1) Construction of the fuzzy comparison matrix

a) Define the CSFs

Note the CSFs for the phase:

$$CSF_1, CSF_2, \ldots, CSF_N$$

(e.g., CSF_1 = Support and involvement of senior management, CSF_2 = Involvement of end users and stakeholders, …)

b) Pairwise comparison (question for consultants): The CSFs are compared in pairs by consultants to assess their relative importance in the phase.

For each pair CSF_i, CSF_j, the following question is asked:

In this phase, is CSF_i more important than CSF_j? And to what extent?

c) Conversion to a triangular fuzzy number (TFN)

Each response is converted to a triangular fuzzy number:

$$\tilde{a}_{\{ij\}} = \left(l_{\{ij\}}, m_{\{ij\}}, u_{\{ij\}}\right)$$

Interpretation:

- l = lowest possible value
- m = most likely value
- u = highest possible value

Linguistic judgments are converted into fuzzy numbers according to the scale shown in the following Table IV.

d) Aggregation (average) when there are multiple consultants.

When multiple consultants participate in the evaluation, and each provides (l_k, m_k, u_k), then the final cell is:

$$l_{\{ij\}} = \frac{1}{K}\sum_{k=1}^{k} l_{i,j}^{(K)}$$

$$m_{\{ij\}} = \frac{1}{K}\sum_{k=1}^{k} m_{i,j}^{(K)}$$

$$u_{\{ij\}} = \frac{1}{K}\sum_{k=1}^{k} u_{i,j}^{(K)}$$

TABLE IV. CONVERTING LINGUISTIC JUDGMENTS INTO TRIANGULAR FUZZY NUMBERS.

Linguistic variables	Fuzzy scale	Corresponding fuzzy scale
Equally important	(1, 1, 1)	(1, 1, 1)
Moderately important	(1/2, 1, 3/2)	(2/3, 1, 2)
Important	(3/2, 2, 5/2)	(2/5, 1/2, 2/3)
Very important	(5/2, 3, 7/2)	(2/7, 1/3, 2/5)
Extremely important	(7/2, 4, 9/2)	(2/9, 1/4, 2/7)

where, K represents the number of consultants.

Therefore:

$$\tilde{a}_{\{ij\}} = \left(\frac{1}{K}\sum_{k=1}^{k} l_{i,j}^{(K)}, \frac{1}{K}\sum_{k=1}^{k} m_{i,j}^{(K)}, \frac{1}{K}\sum_{k=1}^{k} u_{i,j}^{(K)}\right)$$

e) Consistency rules for the matrix

Diagonal:

A criterion compared to itself is equal to:

$$\tilde{a}_{\{ii\}} = (1,1,1)$$

Reciprocal (inverse):

If:

$$\tilde{a}_{\{ij\}} = (l, m, u)$$

So:

$$\tilde{a}_{\{ji\}} = \tilde{a}_{\{ij\}}^{\{-1\}} = \left(\frac{1}{u}, \frac{1}{m}, \frac{1}{l}\right)$$

f) Result of this step: We obtain the fuzzy comparison matrix:

$$\tilde{A} = \begin{bmatrix} 1,1,1 & \tilde{a}_{\{12\}} & \tilde{a}_{\{13\}} & \tilde{a}_{\{14\}} & \tilde{a}_{\{15\}} & \tilde{a}_{\{16\}} \\ \tilde{a}_{\{21\}} & 1,1,1 & \tilde{a}_{\{23\}} & \tilde{a}_{\{24\}} & \tilde{a}_{\{25\}} & \tilde{a}_{\{25\}} \\ \tilde{a}_{\{31\}} & \tilde{a}_{\{32\}} & 1,1,1 & \tilde{a}_{\{34\}} & \tilde{a}_{\{35\}} & \tilde{a}_{\{36\}} \\ \tilde{a}_{\{41\}} & \tilde{a}_{\{42\}} & \tilde{a}_{\{43\}} & 1,1,1 & \tilde{a}_{\{45\}} & \tilde{a}_{\{46\}} \\ \tilde{a}_{\{51\}} & \tilde{a}_{\{52\}} & \tilde{a}_{\{53\}} & \tilde{a}_{\{54\}} & 1,1,1 & \tilde{a}_{\{56\}} \\ \tilde{a}_{\{61\}} & \tilde{a}_{\{62\}} & \tilde{a}_{\{63\}} & \tilde{a}_{\{64\}} & \tilde{a}_{\{65\}} & 1,1,1 \end{bmatrix}$$

2) Calculation of fuzzy sums by criterion: For each CSF, the sum of the fuzzy values in the corresponding row is calculated:

$$\Sigma l_i = \sum_{j=1}^{n} l_{i,j}$$

$$\Sigma m_i = \sum_{j=1}^{n} m_{i,j}$$

$$\Sigma u_i = \sum_{j=1}^{n} u_{i,j}$$

These values represent the cumulative weight of a CSF relative to the others.

The total sums are then calculated:

$$\Sigma l_{tot} = \sum_{i=1}^{n} \Sigma l_i$$

$$\Sigma m_{tot} = \sum_{i=1}^{n} \Sigma m_i$$

$$\Sigma u_{tot} = \sum_{i=1}^{n} \Sigma u_i$$

3) Calculation of fuzzy synthetic values: Fuzzy synthetic values are obtained through normalization:

$$S_{\{l,i\}} = \Sigma l_i \times \frac{1}{\Sigma u_{tot}}$$

$$S_{\{m,i\}} = \Sigma m_i \times \frac{1}{\Sigma m_{tot}}$$

$$S_{\{u,i\}} = \Sigma u_i \times \frac{1}{\Sigma l_{tot}}$$

These values represent the relative importance of the CSF in a vague form.

4) Defuzzification of weights: To obtain a single usable weight, fuzzy values are converted into simple numerical values using the center-of-gravity method:

$$W_i = \frac{S_{\{l,i\}} + S_{\{m,i\}} + S_{\{u,i\}}}{3}$$

This step provides a numerical value representing the significance of the CSF.

5) Normalization of weights: The weights obtained are then normalized so that their sum equals 1:

$$w_i = \frac{W_i}{\sum_{k=1}^{n} W_k}$$

The weights can then be expressed as a percentage:

$$Weight_i(\%) = 100 * w_i$$

These values correspond to the final relative importance of the CSF in the phase.

V. FINDINGS AND RESULTS

A. Identification of the CSFs for Each Phase of the ERPI

As explained in the literature review, the ERPI process is divided into three main stages, namely: pre-implementation, implementation and post-implementation.

In our research, pre-implementation includes the analysis phase and the design phase. While implementation includes the realization phase and the final preparation phase. The last step is the post-implementation which includes the commissioning and support phase. Table V details the CSFs for each phase of the ERPI.

1) Analysis phase: According to [56], the analysis phase is a key step in the ERPI, as it aims to establish a thorough understanding of the company's needs. This phase allows the scope of the project, the strategic objectives, and the terms of collaboration to be defined. It involves organizing the project kick-off meeting, structuring the project team, analyzing existing documents and processes, and conducting interviews with key users. These activities enable the functional requirements of the organization to be identified, formalized, and validated.

2) Design phase: According to [56], the design phase aims to translate the functional requirements validated during the analysis phase into operational solutions adapted to the ERP. It is based primarily on the functional requirements document and consists of designing the target processes in detail through the preparation and presentation of prototypes of the various ERP modules to be implemented. Workshops are organized with stakeholders to define management and customization rules and any additional features. This phase also allows the data migration strategy and methods to be defined.

3) Realization phase: According to [56], the realization phase involves the practical deployment of the solutions defined during the design phase. This phase includes configuring the various ERP modules to be implemented, carrying out customizations and specific developments, as well as preparing and migrating data to the test environment. Tests are then conducted to verify that the system meets functional and technical requirements. Any detected issues are analyzed and corrected iteratively to ensure the stability and quality of the solution.

4) Final preparation phase: According to [56], the final preparation phase marks the transition period between the technical implementation of the ERP system and its actual deployment. This phase involves organizing training sessions for users to ensure they are able to use the system effectively. It also includes conducting user acceptance testing (UAT) to verify that the ERP system meets functional and operational requirements. Furthermore, any issues detected during UAT are analyzed and corrected, thereby ensuring the system's quality and reliability before it goes live.

5) The commissioning and support phase: According to [56], the commissioning and support phase is the final stage of the ERPP, where the system is deployed in the production environment. This phase is based on the final configuration of the system, the source code for specific and customized developments, the data ready for migration, and the production plan established during the previous phase. The main activities include configuring the production environment, installing the source code, migrating data to the production environment, conducting a production audit to verify system compliance and performance, and setting up post-production support to assist users and resolve any incidents.

Using the action research method, we identified the CSFs that determine the success of ERPI within Moroccan SMEs. In total, we identified nine factors that influence the success of ERPI in this context. Below, we present the main CSFs identified in this study, as well as the phases during which each of these factors exerts an influence throughout the ERPI process.

Management support and involvement are essential to the successful ERPI. This depends on the willingness of executives to provide sufficient and ongoing support for the system implementation process. Management must be fully involved in decision-making, as an ERPP entails significant changes to existing business processes. Management's support for employees is a prerequisite for the acceptance of the new system. Indeed, if management demonstrates commitment to the project, employees will be more inclined to adopt it. Conversely, a lack of buy-in from management can hinder user acceptance. It is therefore crucial that management does not express doubts and that it demonstrates confidence in the system's value. It must also foster an environment conducive to the ERP project's acceptance; otherwise, the project risks being rejected. This factor remains essential throughout all phases of the ERPI.

Involvement of end users and stakeholders is essential to a project's success. Their participation in the implementation process helps to better understand requirements and needs, validate business rules, customization needs, and new features to be developed, as well as test the system and participate in training sessions. Insufficient user participation often leads to project failure, as a lack of representation can result in ill-suited business processes. This factor is essential in the various phases of ERPI, with the exception of the development phase.

Project management is essential to the successful implementation of the ERP. It consists of continuously managing the implementation plan. This includes planning the various stages, determining milestones, allocating responsibilities and defining success criteria. Effective project management ensures that the core activities of the ERPI are given priority and helps to meet deadlines and timelines. This factor is necessary in all phases of the ERPI.

Process reengineering is the process of observing and analyzing how the business operates to determine what changes can be made to streamline its operations. It is a way of examining the current operations of the business to determine the best way to proceed when designing the new system. This factor is necessary in the analysis phase.

Project team competence is essential for ensuring the success of the implementation. The team must possess the skills, knowledge, influence, and experience necessary to carry out the project, in order to effectively manage meetings with users, ask the right questions to define requirements and lead process reengineering, prepare and present the prototype to users, validate with them the management rules and

customization needs for each module, and define the framework for data migration. It is also crucial to identify experts with a deep understanding of the company, its operations, and its objectives to ensure the project's success. This factor is essential throughout all phases of the ERPI.

Change management is essential to the successful ERPI. To address any resistance you may encounter, it is important to help employees understand the benefits of digital transformation that result from the deployment of the new system. Implementing such software within an organization requires a well-thought-out change management process to ensure that the entire company is moving in the same direction. This factor is essential throughout all phases of the ERPI.

Minimal customization is required for the successful ERPI. Customizing an ERP can be really interesting and beneficial, but it can also result in additional costs and an even longer implementation timeline, so it becomes difficult to make upgrades to the latest versions. This factor is necessary in the realization phase.

Configuration ERP deals with the management of numerous usage controls, which can be enabled or disabled, to balance its functionality to the extent necessary. The configuration of these elements is based on the detailed functional specifications to align the ERPS processes with the business processes of the company. The configuration must be performed by people who understand the business and the company's business, and also have a thorough understanding of the solution to be implemented. This factor is necessary in the realization phase.

User training is essential to ensure that the new system is well integrated internally and used correctly. It plays an essential role in improving the performance of the organization. This factor is necessary in the final preparation phase.

Tests are essential for to allow users to verify that the ERP system is functioning properly before it is deployed in the production environment. These tests help identify and resolve any issues that are detected, thereby ensuring the system's reliability when it goes live.

TABLE V. CSFS FOR EACH PHASE OF THE ERPI

CSFs	Analysis	Design	Realization	Final preparation	Commissioning and support
Management support and involvement	✓	✓	✓	✓	✓
Involvement of end users and stakeholders	✓	✓	—	✓	✓
Project management	✓	✓	✓	✓	✓
Process reengineering	✓	—	—	—	—
Project team competence	✓	✓	✓	✓	✓
Change management	✓	✓	✓	✓	✓
Minimal customization	—	—	✓	—	—
Configuration ERP	—	—	✓	—	—
User training	—	—	—	✓	—
Tests	—	—	—	✓	—

B. Priorisation of the CSFs for Each Phase of the ERPI

1) Analysis phase: In the decision-making process, which will utilize the fuzzy AHP method for the analysis phase, the CSFs are coded as CSF1, CSF2, CSF3, CSF4, CSF5 and CSF6. These CSFs and their corresponding codes are presented below (see Table VI):

CSF1: Management support and involvement

CSF2: Involvement of end users and stakeholders

CSF3: Project management

CSF4: Process reengineering

CSF5: Project team competence

CSF6: Change management

TABLE VI. FUZZY PAIRWISE COMPARISON MATRIX OF CRITICAL SUCCESS FACTORS IN THE ANALYSIS PHASE

	CSF1			CSF2			CSF3			CSF4			CSF5			CSF6		
	l	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**
CSF1	1	1	1	1,305	1,417	1,529	1,139	1,238	1,337	1,218	1,324	1,431	1,774	1,929	2,083	2,258	2,455	2,651
CSF2	0,654	0,706	0,766	1	1	1	0,873	0,944	1,015	0,933	1,01	1,087	1,357	1,476	1,595	1,728	1,88	2,032
CSF3	0,748	0,808	0,878	0,985	1,059	1,145	1	1	1	1,07	1,157	1,244	1,558	1,694	1,83	1,985	2,158	2,33
CSF4	0,699	0,755	0,821	0,92	0,99	1,072	0,804	0,864	0,935	1	1	1	1,455	1,582	1,709	1,854	2,015	2,176
CSF5	0,48	0,518	0,564	0,627	0,677	0,731	0,546	0,59	0,639	0,585	0,632	0,684	1	1	1	1,274	1,386	1,497
CSF6	0,377	0,407	0,443	0,492	0,532	0,575	0,429	0,463	0,501	0,459	0,496	0,537	0,668	0,721	0,78	1	1	1

TABLE VII. FUZZY SYNTHETIC EXTENT VALUES AND RELATIVE IMPORTANCE WEIGHTS OF CRITICAL SUCCESS FACTORS IN THE ANALYSIS PHASE

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
CSF1	8,694	9,363	10,031	0,023	0,025	0,027	0,2	0,234	0,271	0,23	23
CSF2	6,545	7,016	7,495	0,023	0,025	0,027	0,151	0,175	0,202	0,18	18
CSF3	7,346	7,876	8,427	0,023	0,025	0,027	0,169	0,197	0,228	0,2	20
CSF4	6,732	7,206	7,713	0,023	0,025	0,027	0,155	0,18	0,208	0,18	18
CSF5	4,512	4,803	5,115	0,023	0,025	0,027	0,104	0,12	0,138	0,12	12
CSF6	3,425	3,619	3,836	0,023	0,025	0,027	0,079	0,09	0,104	0,09	9
Sum	37,254	39,883	42,617								

The results show that CSF1 is the most important CSF during this phase, accounting for 23% of the total. This underscores the importance of management's commitment to decision-making, allocating the necessary resources, resolving issues, and ensuring clear communication and employee support to foster acceptance of the new ERP system (see Table VII).

The second most important factor in this phase is CSF3, which accounts for 20% of the total. Effective project management is essential to ensure that all activities in this phase are properly planned and executed on time, in order to meet the project's objectives. This is why it is necessary to have a competent project manager who can assign responsibilities, manage resources, and ensure clear communication with all stakeholders.

Factors CSF2 and CSF4 rank third, each at 18%. User involvement provides a better understanding of the company's processes, specific needs, and the issues and challenges it faces. Furthermore, process reengineering helps identify and optimize business processes in order to align them with the requirements of the ERP system.

Next, CSF5 accounts for 12%, indicating that the project team's competence remains essential for effectively managing meetings with users, asking the right questions to define requirements, and reengineering processes.

Finally, CSF6 has the lowest weight at 9%. This explains why activities related to change management tend to become more important during the later phases of ERPI.

2) Design phase: In the decision-making process, which will utilize the fuzzy AHP method for the design phase, the CSFs are coded as CSF1, CSF2, CSF3, CSF4, and CSF5. These CSFs and their corresponding codes are presented below (see Table VIII):

CSF1: Management support and involvement

CSF2: Involvement of end users and stakeholders

CSF3: Project management

CSF4: Project team competence

CSF5: Change management

TABLE VIII. FUZZY PAIRWISE COMPARISON MATRIX OF CRITICAL SUCCESS FACTORS IN THE DESIGN PHASE

	CSF1			CSF2			CSF3			CSF4			CSF5		
	l	m	u	l	m	u	l	m	u	l	m	u	L	m	u
CSF1	1	1	1	0,846	0,917	0,987	0,599	0,65	0,702	0,735	0,798	0,861	1,057	1,146	1,234
CSF2	1,013	1,091	1,182	1	1	1	0,708	0,769	0,83	0,869	0,944	1,018	1,249	1,354	1,458
CSF3	1,425	1,538	1,67	1,205	1,3	1,412	1	1	1	1,228	1,333	1,438	1,765	1,913	2,06
CSF4	1,161	1,253	1,361	0,982	1,059	1,151	0,695	0,75	0,814	1	1	1	1,438	1,559	1,679
CSF5	0,81	0,873	0,946	0,686	0,738	0,801	0,485	0,523	0,567	0,596	0,641	0,695	1	1	1

TABLE IX. FUZZY SYNTHETIC EXTENT VALUES AND RELATIVE IMPORTANCE WEIGHTS OF CRITICAL SUCCESS FACTORS IN THE DESIGN PHASE

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
CSF1	4,237	4,511	4,784	0,036	0,038	0,041	0,153	0,171	0,196	0,17	17
CSF2	4,839	5,158	5,488	0,036	0,038	0,041	0,174	0,196	0,225	0,2	20
CSF3	6,623	7,084	7,58	0,036	0,038	0,041	0,238	0,269	0,311	0,27	27
CSF4	5,276	5,621	6,005	0,036	0,038	0,041	0,19	0,214	0,246	0,22	22
CSF5	3,577	3,775	4,009	0,036	0,038	0,041	0,129	0,143	0,164	0,14	14
Sum	24,552	26,149	27,866								

The results show that CSF3 is the most important CSF during this phase, accounting for 27% of the total. Effective project management is essential to ensure that all activities in this phase are properly planned and executed on time, in order to meet the project's objectives. This is why it is necessary to have a competent project manager who can assign responsibilities, manage resources, and ensure clear communication with all stakeholders.

The second most important factor is CSF4, which accounts for 22% of the total. Project team competence is essential for effectively preparing and presenting the prototype to users, facilitating meetings with them, asking the right questions to identify the management rules and customization needs for each module, and defining the framework for data migration.

Next, the CSF2 accounts for 20%. The involvement of end users and stakeholders is essential for understanding requirements, needs, issues, and challenges, as well as for validating business rules, customization needs, and new features to be developed. It also helps ensure that the ERP effectively meets the organization's actual needs.

CSF1 accounts for 17%. Although management support and involvement are essential for making decisions, managing resources, resolving issues, ensuring clear communication, and supporting employees to foster acceptance of the new ERP system, their influence is slightly less decisive in this phase compared to other factors.

Finally, CSF5 accounts for 14% of the weight. The importance of change management begins to grow during the final phases of the project in order to help prepare employees for the changes brought about by the implementation of the ERP system and to minimize resistance (see Table IX).

3) Realization phase: In this decision-making process, which will utilize the fuzzy AHP method for the realization phase, the CSFs are coded as CSF1, CSF2, CSF3, CSF4, CSF5, and CSF6. These CSFs and their corresponding codes are presented below (see Table X):

CSF1: Management support and involvement

CSF2: Project management

CSF3: Project team competence

CSF4: Change management

CSF5: Minimal customization

CSF6: Configuration

The results show that CSF6 is the most important CSF in this phase, accounting for 24% of the total. Proper ERP configuration is essential for aligning business processes with the ERP in order to meet the organization's needs and requirements (see Table XI).

TABLE X. FUZZY PAIRWISE COMPARISON MATRIX OF CRITICAL SUCCESS FACTORS IN THE REALIZATION PHASE

	CSF1			CSF2			CSF3			CSF4			CSF5			CSF6		
	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u
CSF1	1	1	1	0,394	0,429	0,463	0,436	0,474	0,512	0,69	0,75	0,81	0,552	0,6	0,648	0,345	0,375	0,405
CSF2	2,16	2,333	2,536	1	1	1	1,017	1,105	1,194	1,61	1,75	1,89	1,288	1,4	1,512	0,805	0,875	0,945
CSF3	1,955	2,111	2,295	0,838	0,905	0,983	1	1	1	1,457	1,583	1,71	1,165	1,267	1,368	0,728	0,792	0,855
CSF4	1,235	1,333	1,449	0,529	0,571	0,621	0,585	0,632	0,686	1	1	1	0,736	0,8	0,864	0,46	0,5	0,54
CSF5	1,543	1,667	1,812	0,661	0,714	0,776	0,731	0,789	0,858	1,157	1,25	1,359	1	1	1	0,575	0,625	0,675
CSF6	2,469	2,667	2,899	1,058	1,143	1,242	1,17	1,263	1,373	1,852	2	2,174	1,481	1,6	1,739	1	1	1

TABLE XI. FUZZY SYNTHETIC EXTENT VALUES AND RELATIVE IMPORTANCE WEIGHTS OF CRITICAL SUCCESS FACTORS IN THE REALIZATION PHASE

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
CSF1	3,417	3,628	3,838	0,023	0,025	0,027	0,079	0,091	0,104	0,09	9
CSF2	7,88	8,463	9,077	0,023	0,025	0,027	0,181	0,212	0,245	0,21	21
CSF3	7,143	7,658	8,211	0,023	0,025	0,027	0,164	0,191	0,222	0,19	19
CSF4	4,545	4,836	5,16	0,023	0,025	0,027	0,105	0,121	0,139	0,12	12
CSF5	5,667	6,045	6,48	0,023	0,025	0,027	0,13	0,151	0,175	0,15	15
CSF6	9,03	9,673	10,427	0,023	0,025	0,027	0,208	0,242	0,282	0,24	24
Sum	37,682	40,303	43,193								

The second most important factor is CSF2, accounting for 21%. Effective project management is essential to ensure that all activities in this phase are properly planned and executed on time, in order to meet the project's objectives. This is why it is necessary to have a competent project manager who can assign responsibilities, manage resources, and ensure clear communication with all stakeholders.

Next, CSF3 accounts for 19%. Project team competence is essential to properly configure and customize the ERP, as well

as to migrate data, perform tests, and resolve any issues detected during those tests.

The CSF5 is assigned a weight of 15%. Minimal customization is essential to avoid complicating upgrades to new versions of the ERP system, thereby reducing costs and simplifying future upgrades.

Furthermore, CSF4 is assigned a weight of 12%. Although change management is essential for preparing users and helping them adopt the new system, its importance generally becomes more critical in later phases.

Finally, CSF1 receives the lowest weighting, 9% in this phase. Although management support and involvement are essential for making decisions, managing resources, resolving issues, ensuring clear communication, and supporting employees to foster acceptance of the new ERP system, their influence is slightly less decisive in this phase compared to other factors.

4) Final preparation phase: In the decision-making process, which will utilize the fuzzy AHP method for the realization phase, the CSFs are coded as CSF1, CSF2, CSF3, CSF4, CSF5, CSF6, and CSF7. These CSFs and their corresponding codes are presented below (see Table XII):

CSF1: Management support and involvement

CSF2: Involvement of end users and stakeholders

CSF3: Project management

CSF4: Project team competence

CSF5: Change management

CSF6: User training

CSF7: Tests

TABLE XII. FUZZY PAIRWISE COMPARISON MATRIX OF CRITICAL SUCCESS FACTORS IN THE FINAL PREPARATION PHASE

	CSF1			CSF2			CSF3			CSF4			CSF5			CSF6			CSF7		
	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u
CSF 1	1	1	1	0,328	0,357	0,386	0,41	0,446	0,482	0,615	0,669	0,723	0,273	0,297	0,321	0,223	0,243	0,263	0,246	0,268	0,29
CSF 2	2,591	2,8	3,049	1	1	1	1,25	1,357	1,464	1,875	2,036	2,196	0,833	0,905	0,976	0,682	0,741	0,8	0,75	0,815	0,879
CSF 3	2,074	2,241	2,439	0,683	0,737	0,8	1	1	1	1,5	1,629	1,757	0,667	0,724	0,781	0,545	0,592	0,638	0,6	0,652	0,703
CSF 4	1,383	1,495	1,626	0,455	0,491	0,533	0,569	0,614	0,667	1	1	1	0,379	0,412	0,444	0,31	0,337	0,364	0,341	0,37	0,4
CSF 5	3,115	3,367	3,663	1,025	1,105	1,2	1,28	1,381	1,499	2,25	2,429	2,637	1	1	1	0,818	0,889	0,96	0,9	0,978	1,056
CSF 6	3,802	4,111	4,474	1,25	1,35	1,466	1,567	1,688	1,835	2,75	2,964	3,214	1,042	1,125	1,222	1	1	1	1,1	1,196	1,291
CSF 7	3,447	3,731	4,065	1,138	1,227	1,333	1,423	1,534	1,667	2,5	2,703	2,933	0,947	1,022	1,111	0,775	0,836	0,909	1	1	1

TABLE XIII. FUZZY SYNTHETIC EXTENT VALUES AND RELATIVE IMPORTANCE WEIGHTS OF CRITICAL SUCCESS FACTORS IN THE FINAL PREPARATION PHASE

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
CSF1	2,849	3,012	3,175	0,016	0,018	0,019	0,046	0,054	0,06	0,05	5
CSF2	8,231	8,839	9,485	0,016	0,018	0,019	0,132	0,159	0,18	0,15	15
CSF3	6,469	6,923	7,415	0,016	0,018	0,019	0,104	0,125	0,141	0,12	12
CSF4	4,096	4,349	4,634	0,016	0,018	0,019	0,066	0,078	0,088	0,08	8
CSF5	9,488	10,171	10,959	0,016	0,018	0,019	0,152	0,183	0,208	0,18	18
CSF6	11,411	12,238	13,211	0,016	0,018	0,019	0,183	0,22	0,251	0,22	22
CSF7	10,23	11,053	12,018	0,016	0,018	0,019	0,164	0,199	0,228	0,2	20
Sum	52,774	56,585	60,897								

The results indicate that CSF6 is the most important CSF in this phase, with a weight of 22%. User training is essential, as this phase precedes the implementation of the ERP system in the production environment. It is therefore important to thoroughly train users on how to use the new ERP system, which helps reduce errors during use and also facilitates its adoption (see Table XIII).

The second most important factor CSF7, accounting for 20%. Testing is essential to allow users to test and verify that the ERP is functioning properly before it is deployed in the production environment. These tests help identify and correct any issues that are detected, ensuring the system's reliability when it goes live.

Next, CSF5 accounts for 18%. Change management is becoming increasingly important in helping to prepare employees for the changes brought about by the implementation of the ERP system, as well as in minimizing resistance to change.

The CSF2 is weighted at 15%. The involvement of end users and stakeholders is essential during training sessions and testing to ensure that the new ERP system meets the organization's needs.

Furthermore, CSF3 accounts for 12%. Effective project management is essential to ensure that all activities in this phase are properly planned and executed on time, in order to meet the project's objectives. This is why it is necessary to have a competent project manager who can assign responsibilities, manage resources, and ensure clear communication with all stakeholders.

Next, CSF4 accounts for 8%. Project team competence is essential for effectively training users, conducting user testing, and resolving issues identified during testing.

Finally, CSF1 receives the lowest weighting, 5%, in this phase. Although management support and involvement are essential for making decisions, managing resources, resolving issues, ensuring clear communication, and supporting employees to foster acceptance of the new ERP system, their influence is slightly less decisive in this phase compared to other factors.

5) Commissioning and support phase: In the decision-making process, which will utilize the fuzzy AHP method for the commissioning and support phase, the CSFs are coded as CSF1, CSF2, CSF3, CSF4, and CSF5. These CSFs and their corresponding codes are presented below (see Table XIV).

CSF1: Management support and involvement

CSF2: Involvement of end users and stakeholders

CSF3: Project management

CSF4: Project team competence

CSF5: Change management

The results indicate that CSF5 is the most important CSF in this phase, accounting for 27% of the total. During this phase, employees begin to use the new ERP, which entails changes to the organization's business processes. Change management is essential during this phase to help employees adapt to the changes brought about by the implementation of the ERP and to minimize resistance to change (see Table XV).

TABLE XIV. FUZZY PAIRWISE COMPARISON MATRIX OF CRITICAL SUCCESS FACTORS IN THE COMMISSIONING AND SUPPORT PHASE

	CSF1			CSF2			CSF3			CSF4			CSF5		
	l	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**	**l**	**m**	**u**
CSF1	1	1	1	0,56	0,608	0,656	0,723	0,786	0,849	0,647	0,703	0,759	0,46	0,5	0,54
CSF2	1,524	1,645	1,786	1	1	1	1,312	1,429	1,545	1,174	1,278	1,382	0,746	0,811	0,876
CSF3	1,178	1,273	1,383	0,647	0,7	0,762	1	1	1	0,891	0,969	1,047	0,566	0,615	0,664
CSF4	1,317	1,423	1,545	0,724	0,783	0,852	0,955	1,032	1,122	1	1	1	0,635	0,69	0,745
CSF5	1,852	2	2,174	1,142	1,233	1,339	1,506	1,625	1,758	1,342	1,457	1,576	1	1	1

TABLE XV. FUZZY SYNTHETIC EXTENT VALUES AND RELATIVE IMPORTANCE WEIGHTS OF CRITICAL SUCCESS FACTORS IN THE COMMISSIONING AND SUPPORT PHASE.

	Σl	**Σm**	**Σu**	**1/Σu**	**1/Σm**	**1/Σl**	**S_l**	**S_m**	**S_u**	**W**	**%**
CSF1	3,39	3,597	3,804	0,035	0,038	0,04	0,119	0,137	0,152	0,14	14
CSF2	5,756	6,163	6,589	0,035	0,038	0,04	0,201	0,234	0,264	0,23	23
CSF3	4,282	4,557	4,856	0,035	0,038	0,04	0,15	0,173	0,194	0,17	17
CSF4	4,631	4,928	5,264	0,035	0,038	0,04	0,162	0,187	0,211	0,19	19
CSF5	6,842	7,315	7,847	0,035	0,038	0,04	0,239	0,278	0,314	0,27	27
Sum	24,901	26,56	28,36								

The second most important factor is CSF2, accounting for 23%. The involvement of end users and stakeholders is essential for getting the system up and running, as it allows for the rapid identification of operational issues and anomalies, thereby contributing to the system's continuous improvement after implementation.

Next, the CSF4 accounts for 19%. Project team competence is essential for properly configuring the production environment, implementing customizations, migrating data, assisting users, and resolving any issues that arise.

The CSF3 accounts for 17%. Effective project management is essential to ensure that all activities in this phase are properly planned and executed on time, in order to meet the project's objectives. This is why it is necessary to have a competent project manager who can assign responsibilities, manage resources, and ensure clear communication with all stakeholders.

Finally, CSF1 account for 14% of the total. Although management support and involvement are essential for making decisions, managing resources, resolving issues, ensuring clear communication, and supporting employees to foster acceptance

of the new ERP system, their influence is slightly less decisive in this phase compared to other factors.

C. Changes in CSFs Across ERPI Phases

To better understand the dynamics of ERPPs, CSFs have been classified into four main categories: strategic factors, organizational factors, project management factors, and technological factors. Analysis of the results shows that the relative importance of these categories varies significantly across the different phases of ERPI (see Table XVI).

During the analysis phase, strategic factors emerge as the most critical. Management support and involvement are the most important factors, as they enable the project's direction to be defined, the necessary resources to be mobilized, and the alignment of the ERP with the company's overall strategy to be ensured. Factors related to project management and process reengineering also play a key role in the initial structuring of the project.

During the design phase, organizational factors and factors related to project management take on greater importance. Project management and the project team competence are central to this phase, as they enable the organizational needs identified during the analysis phase to be transformed into appropriate functional solutions.

During the implementation phase, technological factors take center stage. The configuration ERP is the primary critical factor, followed by project management and the project team's competence. This phase is characterized by technical activities such as configuration and development.

During the final preparation phase, the importance of organizational factors becomes increasingly evident through user training and testing activities. These factors help prepare the organization for the actual use of the ERP and reduce the risks associated with the go-live.

Finally, during the implementation and support phase, organizational factors related to change management and end-user engagement become critical. These factors help foster acceptance of the system, support users as they adapt to new processes, and ensure the continued operation of the ERP.

The study's findings show that CSFs do not carry the same weight throughout the ERPP. Their priority shifts according to the specific requirements of each implementation phase. Strategic factors dominate the analysis phase, while project management and technological factors become more important during the design and implementation phases. Finally, organizational factors take on major importance during the final preparation and system deployment phases.

TABLE XVI. CHANGES IN CSFS ACROSS ERPI PHASES

	Analysis	Design	Realization	Final preparation	Commissioning and support
Management support and involvement	23 %	17 %	9 %	5 %	14 %
Involvement of end users and stakeholders	18 %	20 %	—	15 %	23 %
Project management	20 %	27 %	21 %	12 %	17 %
Process reengineering	18 %	—	—	—	—
Project team competence	12 %	22 %	19 %	8 %	19 %
Change management	9 %	14 %	12 %	18 %	27 %
Minimal customization	—	—	15 %	—	—
Configuration	—	—	24 %	—	—
User training	—	—	—	22 %	—
Tests	—	—	—	20 %	—

VI. DISCUSSION

To provide a deeper understanding of the results obtained, this section discusses them in relation to previous research on the CSFs of ERPI projects. The analysis highlights both areas of agreement with the existing literature and specific contributions related to the phased approach adopted in this study.

First, the results confirm the importance of strategic factors during the ERPP analysis phase. In particular, the support and involvement of senior management emerge as a determining factor for the project's success. This finding aligns with the conclusions of several previous studies that emphasize the central role of management commitment in defining the project vision, allocating the necessary resources, and aligning the ERP with the company's overall strategy [19,28,46].

Furthermore, the results show that factors related to project management and technological factors become increasingly important during the design and implementation phases. Efficient project management, the project team competence, and the configuration of the ERP emerge as essential elements for transforming organizational needs into appropriate functional and technical solutions. These conclusions are consistent with the studies by [18,33,38], who highlight the key role of project planning, the team's technical skills, and proper system configuration in the success of ERPI.

Furthermore, the study highlights the growing importance of organizational factors in the final phases of the project, particularly during the final preparation phase and the implementation and support phase. User training, change management, and end-user involvement appear to be essential conditions for promoting system adoption and ensuring its continued use. These results confirm the findings of several

studies emphasizing that the success of an ERPP depends not only on technical aspects but also on the organization's ability to support the organizational transformations brought about by the system's implementation [47,48].

Overall, the main contribution of this research lies in highlighting how CSFs evolve over the course of an ERPP lifecycle. Unlike many studies that analyze CSFs in a general manner, this research shows that their relative importance varies across the different phases of the implementation process. This phase-based approach provides a more nuanced understanding of the conditions for ERPP success and offers practical guidance to decision-makers and consultants involved in the implementation of ERP within Moroccan SMEs.

VII. Research Contributions and Limitations

This work makes several contributions, both theoretical and practical.

From a theoretical standpoint, this research:

- Complements the existing literature, which generally analyzes CSFs in a broad sense, by introducing a phase-based analysis of the ERPP.
- Offers a dynamic approach to CSFs by demonstrating that their relative importance changes across the different phases of the ERPI process.
- Contributes to an improved understanding of the ERPI process by highlighting the specific role of strategic, organizational, project management, and technological factors at each phase of the project.

From a practical standpoint, this research:

- Provides ERP consultants and project managers with a clear overview of the factors to prioritize at each stage of the project.
- Helps ERP consultants and project managers better tailor their support strategies to the specific requirements of each implementation phase.
- Contributes to reducing the risk of ERPP failure by identifying the CSFs that ensure a successful system deployment.

Our research has certain limitations:

- It was limited to Moroccan SMEs.
- The action research method is used in only two enterprises.

VIII. Conclusion and Future Works

The objective of this study is to identify the CSFs associated with each phase of the ERPI process, as well as to determine their relative importance in each of these phases within Moroccan SMEs.

In this research, we adopted a mixed-methods approach. The action research method was used to identify the CSFs associated with each phase of the ERPI process within Moroccan SMEs. Additionally, we used the Fuzzy AHP method to determine their relative importance in each of these phases.

The results indicate that the importance of CSFs varies across the different phases of ERPI. During the analysis phase, strategic factors are paramount, while those related to project management and technology gain importance during the design and implementation phases. Finally, organizational factors play a crucial role during the final preparation phase, as well as during the implementation and support phase.

This work provides numerous themes for future research:

- Expand this study to include a larger number of Moroccan SMEs.
- Expand this study to include enterprises of other sizes and other countries.
- Examine the importance of CSFs after the ERPI.
- Examine the challenges that hinder the success of ERPI in the context of Moroccan SMEs.

References

[1] Al-Mashari, M., Zairi, M., 2000. Supply-chain re-engineering using enterprise resource planning

[2] EL MRINI, Y., BELAISSAOUI, M., & TAQAFI, I. Difficultés liées à la mise en œuvre des ERPs au Maroc: Focus sur Gouvernance.

[3] Malhotra, R. and Temponi, C., 2010. Critical decisions for ERP integration: small business issues. International Journal of Information Management, 30 (1), 28–37.

[4] Upadhyay, P., Jahanyan, S., and Dan, P.K., 2011. Factors influencing ERP implementation in Indian manufacturing organisations: a study of micro, small and medium-scale enterprises. Journal of Enterprise Information Management, 24 (2), 130–145.

[5] Olson, D.L. and Staley, J., 2012. Case study of open-source enterprise resource planning implementation in a small business. Enterprise Information Systems, 6 (1), 79–94.

[6] A. Jones, J. Robinson, B. O'Toole, and D. Webb, "Implementing a bespoke supply chain management system to deliver tangible benefits," International Journal of Advanced Manufacturing Technology, vol. 30, no. 9/10, pp. 927–937, 2006, DOI: 10.1007/s00170-005-0065-2.

[7] Ngai, E. W. T., Cheng, T. C. E., & Ho, S. S. (2004). Critical success factors of web-based supply-chain management systems: an exploratory study. Production Planning & Control, 15(6), 622-630.

[8] Markus, M. L., & Robey, D. (1988). Information technology and organizational change: Causal structure in theory and research. Management science, 34(5), 583-598.

[9] Newman, M., & Robey, D. (1992). A social process model of user-analyst relationships. MIS quarterly, 249-266.

[10] Capaldo, G. & Rippa, P. (2009)."A PlannedOriented Approach for ERP Implementation Strategy Selection," Journal of Enterprise Information Management, Vol. 22, No. 6, pp. 642-659.

[11] Esteves, J. M. & Pastor, J. A. (1999)."An ERP life-Cycle-Based Research Agenda," First International Workshop in Enterprise Management and Resource Planning: Methods, Tools and Architectures (EMRPS), Venice, Italy.

[12] Ross, J. W., & Vitale, M. R. (2000). The ERP revolution: surviving vs. thriving. Information systems frontiers, 2(2), 233-241.

[13] Markus, M. L., & Tanis, C. (2000). The enterprise systems experience-from adoption to success. Framing the domains of IT research: Glimpsing the future through the past, 173(2000), 207-173.

[14] Parr, A. & Shanks, G. (2000)."A Model of ERP Project Implementation," Journal of Information Technology.

[15] Shanks, G., Parr, A., Hu, B., Corbitt, B., Thanasankit, T., & Seddon, P. (2000). Differences in critical success factors in ERP systems

implementation in Australia and China: a cultural analysis. ECIS 2000 Proceedings, 53.

[16] Somers, T. M. & Nelson, K. G. (2004)."A Taxonomy of Players and Activities across the ERP Project life Cycle," Information and Management.

[17] Peslak, A. R., Subramanian, G. H., & Clayton, G. E. (2008). The phases of ERP software implementation and maintenance: A model for predicting preferred ERP use. Journal of Computer information systems, 48(2), 25-33.

[18] Yusuf, Y., Gunasekaran, A., & Wu, C. (2006). Implementation of enterprise resource planning in China. Technovation, 26(12), 1324-1336.

[19] Umble, E. J., Haft, R. R. (2003)."Enterprise Resource Planning: implementation Procedures and Critical Success Factors," European Journal of Operational Research.

[20] Mengistie, A. A., Heaton, D. P., & Rainforth, M. (2012). Analysis of the critical success factors for ERP systems implementation in US Federal Ofces. In F. Piazolo & M. [21] Felderer (Eds.), Innovation and future of enterprise information systems (pp. 183–198). Salzburg: Springer

[21] Lin, Y.-C., Lee, M.-H., & Tsernng, H. P. (2003). Construction enterprise resource planning implementation: critical success factors-lesson learning in Taiwan. 20th International Symposium on Automation and Robotics in Construction ISARC, pp 623–628.

[22] Kronbichler, S. A., Ostermann, H., & Staudinger, R. (2009). A review of critical success factors for ERP-projects. The Open Information Systems Journal, 3(1), 14–25.

[23] Anto, J. (2016). An empirical study of enterprise resource planning systems in construction industry. International Research Journal of Engineering and Technology, 3(4), 1310–1315.

[24] Chung, B. Y., Skibniewski, M. J., Lucas, H. C., & Kwak, Y. H. (2008). Analyzing enterprise resource planning system implementation success factors in the engineering-construction industry. Journal of Computing in Civil Engineering, 22(6), 373–382.

[25] Ara, A., & Al-Mudimigh, A. S. (2011). The role and impact of project management in ERP project implementation life cycle. Global Journal of Computer Science and Technology, 11(5), 6–10.

[26] Akkermans, H., & van Helden, K. (2002). Vicious and virtuous cycles in ERP implementation: a case study of interrelations between critical success factors. European Journal of Information Systems, 11(1), 35–46.

[27] Brown, C., & Vessey, I. (1999). ERP implementation approaches: toward a contingency framework. ICIS 1999 Proceedings, 39.

[28] Holland, C. R., & Light, B. (1999). A critical success factors model for ERP implementation. IEEE software, 16(3), 30-36.

[29] Esteves, J. & Pastor, J. (2000)."Toward the Unification of Critical Success Factors for ERP Implementation," Proceedings of the 10th Annual Business Information Technology (BIT) Conference, Manchester, UK.

[30] Roseman, M., Sedera, W. & Gable, G. (2001)."Critical Success Factors of Process Modeling for Enterprise Systems," Proceedings of the Americas Conference on Information Systems, Boston, Massachusetts.

[31] Allen, D., Kern, T. & Havenhand, M. (2002)."ERP Critical Success Factors: An Exploration of the Contextual Factors in Public Sector Institution," Proceedings of the 35th Hawaii International Conference on System Sciences.

[32] Wenrich, K., & Ahmad, N. (2009). Lessons learned during a decade of ERP experience: A case study. International Journal of Enterprise Information Systems (IJEIS), 5(1), 55-73.

[33] Soja, P. (2006). Success factors in ERP systems implementations: lessons from practice. Journal of enterprise information management.

[34] Yusuf, Y., Gunasekaran, A., & Wu, C. (2006). Implementation of enterprise resource planning in China. Technovation, 26(12), 1324-1336.

[35] Umble, E. J., Haft, R. R. & Umble, M. M. (2003)."Enterprise Resource Planning: implementation Procedures and Critical Success Factors," European Journal of Operational Research.

[36] Ali, M. (2017). Developing in-house ERP system for the construction industry in a developing country: a case study. Engineering Management Research, 6(1), 90.

[37] Motwani, J., Akbul, A.Y., & Nidumolu, V. (2005)."Successful Implementation of ERP Systems: a Case Study of an International Automotive Manufacturer," International Journal of Automotive Technology and Management.

[38] Hong, K. K., & Kim, Y. G. (2002). The critical success factors for ERP implementation: an organizational fit perspective. Information & management, 40(1), 25-40.

[39] Kale, P., Banwait, S., & Laroiya, S. (2007). Enterprise resource planning implementation in Indian SMEs: Issues and challenges, pp. 242–248. https://www.csi-sigegov.org/critical_pdf/27_242-248. pdf. Accessed 30 May 2019.

[40] Ahmed, S.M., Ahmad, I., Azhar, S., & Mallikarjuna, S. (2003). Implementation of enterprise resource planning (ERP) systems in the construction industry. In: Construction Research Congress, American Society of Civil Engineers, Reston, VA, pp 1–8.

[41] Tsai, W.- H., Chien, S.- W., Hsu, P.- Y. & Leu, J.- D. (2005)."Identification of Critical Failure Factors in the Implementation of Enterprise Resource Planning (ERP) System in Taiwan's Industries," International Journal of Management and Enterprise Development.

[42] Voordijk, H., Van Leuven, A., & Laan, A. (2003). Enterprise resource planning in a large construction frm: implementation analysis. Construction Management and Economics, 21(5), 511–521.

[43] Chan, C.T.W., & Sin, H.C. (2010). Critical success factors for ERP implementation in Chinese construction companies. In: 2010 8th IEEE international conference on industrial informatics, IEEE, pp 628–633.

[44] Wu, J. H., & Wang, Y. M. (2007). Measuring ERP success: The key-users' viewpoint of the ERP to produce a viable IS in the organization. Computers in Human behavior, 23(3), 1582-1596.

[45] Kumar, V., Maheshwari, B., & Kumar, U. (2003). An investigation of critical management issues in ERP implementation: emperical evidence from Canadian organizations. Technovation, 23(10), 793-807.

[46] Al-Mashari, M., Al-Mudimigh, A. & Zairi, M. (2003)."Enterprise Resource Planning: A Taxonomy of critical factors," European Journal of Operational Research.

[47] Sun, A. Y., Yazdani, A., & Overend, J. D. (2005). Achievement assessment for enterprise resource planning (ERP) system implementations based on critical success factors (CSFs). International journal of production economics, 98(2), 189-203.

[48] Amini, M., & Sadat Safavi, N. (2013). Critical success factors for ERP implementation. International Journal of Information Technology and Information Systems, 5(15), 1–23.

[49] Chung, B. (2007). An analysis of success and failure factors for ERP systems in engineering and construction frms. Doctoral Thesis. University of Maryland.

[50] Shi, J. J., & Halpin, D. W. (2003). Enterprise resource planning for construction business management. Journal of Construction Engineering and Management, 129(2), 214–221.

[51] Bano, F. (2014). Quality of ERP implementation: case study of select indian organizations. International Proceedings of Economics Development and Research, 75(9), 36–39.

[52] Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. MIS quarterly, 369-386.

[53] Soulard, C. T., Compagnone, C., & Lémery, B. (2007). La recherche en partenariat: entre fiction et friction. Natures sciences sociétés, 15(1), 13-22.

[54] Saaty, T. L. (1980). The analytic hierarchy process.

[55] Chang, D. Y. (1996). Applications of the extent analysis method on fuzzy AHP. European journal of operational research, 95(3), 649-655.

[56] ZOUHAIR, Y., EL MRINI, Y., & BELAISSAOUI, M. (2026). Framework for Implementing ERP Integrating Client-Consultant Agency Management within Moroccan SMEs. International Journal of Advanced Computer Science & Applications, 17(3), 838.

Hybrid Data Fusion and Deep Learning for Dynamic Risk Index Modeling in Secure Learning Management Systems

Vani T[1], S. Sathya[2]
Research Scholar, Department of Computer Science and Information Technology, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India[1]
Associate Professor, Department of Computer Science and Information Technology, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India[2]

***Abstract*—The explosive growth of online education platforms has led to increased exposure to cybersecurity threats, which makes secure Learning Management Systems (LMS) a critical requirement. However, the current methods often can't capture user behavior risk and network-level attack patterns at the same time, which causes the threat to be incomplete. This study presents a dynamic cyber risk prediction model by fusing log information of LMS behavior with network intrusion information in the CICIDS2017 dataset. The goal is to create an AI-based model that is able to perform real-time risk assessment using a Dynamic Risk Index (DRI). The methodology includes the combination of feature engineering, hybrid data fusion, machine learning, deep learning (LSTM, DNN), and anomaly detection methods. Experimental results demonstrate that the proposed model achieves an accuracy of 97.6%, an F1-score of 96.9%, and an AUC of 98.5%, outperforming state-of-the-art methods. The robustness and significance of the framework are confirmed by ablation and statistical analyses. The overall study concludes that combining behavioral and network intelligence with dynamic risk scoring improves cyber threat detection and proactive security management in e-learning environments.**

Keywords*—Cybersecurity; Dynamic Risk Index (DRI); e-learning security; behavioral analytics; intrusion detection; hybrid deep learning; LMS logs; CICIDS2017 dataset; anomaly detection; risk prediction

I. Introduction

The rapid digital transformation of education has resulted in the widespread use of online learning platforms. Specifically, online learning platforms called Learning Management Systems (LMS) enable remote education types to assist in assessing, as well as in involving different educators, students, or learners in online learning[1]. While these platforms offer flexibility and accessibility, they also introduce some intense cybersecurity challenges. The growing amount of sensitive information such as student records, academic content, and institutional information makes LMS environments appealing targets in cyberattacks such as phishing, brute-force attacks, and unauthorized access[2]. As online education continues to grow, robust cybersecurity mechanisms must be in place to support institutions around the world.

Traditional cybersecurity approaches in the e-learning environments are mainly based on rule-based systems or network-level intrusion detection mechanisms. While the above methods work well in detecting known attack patterns, they are often poor at detecting subtle behavioral anomalies displayed by users [3]. Conversely, behavioral analytics techniques are focused on the activity patterns of users but may not be aware of the underlying threat at the network level[4]. This separation between behavioral and network perspectives means that not all risks are measured, and limits the effectiveness of the security solutions in place. Some recent research has been carried out into the application of machine learning and deep learning intrusion detection techniques based on datasets such as the CICIDS2017 dataset [5]. However, most of the research is focused on network traffic analysis and fails to integrate real-world LMS behavioural data, making its applicability in the practical setting of e-learning minimal.

The motivation behind this research is linked to the necessity of having a general and dynamic cybersecurity model that involves the incorporation of behavioral and network-based information [6]. As the nature of cyber threats advances, this is no longer a sufficient method of detection[7]. The need to use smart systems that can constantly monitor is increasing, along with demand for systems capable of dynamically evaluating risk and proactively preventing threats. The level of accuracy in detection may be increased, and the views of the overall scope of cyber risks may be taken with the use of LMS log data and benchmarking intrusion detection sets. Moreover, the global introduction of a Dynamic Risk Index (DRI) provides an uninterrupted and descriptive indication of the user risk that enables real-time decision-makers to implement such things as personalized security intervention.

On the foundation of the challenges and the motivations, the following research question is suggested for the study: How can the behavioral analytics data and network intrusion data successfully be integrated so that they become a dynamic and precise cyber risk prediction model in online education settings? The problem with the case is that it does not have a single framework that connects the heterogeneous information sources of user activity to real-time risk scoring. The current solutions are either very fixed classification or inadaptable to a dynamic threat environment.

A. Objectives of the Study

- To propose a hybrid model of cyber risk prediction based on integration of LMS behavioral data and network intrusion data.
- To propose a Dynamic Risk Index (DRI) for real-time user risk assessment as well as detection of anomalies.
- To test the effectiveness of the proposed framework based on advanced machine learning and deep learning techniques.

B. Contributions of the Study

The major contributions of this study are as follows.

- The proposed novel hybrid cyber risk prediction framework is achieved by integrating LMS behavioral analytics with the CICIDS2017 network intrusion data for the complete cybersecurity analysis of e-learning systems.
- To enhance the representation of multi-source threats, fusion strategies at the feature level and decision level are proposed for fusing behavioral log and network traffic features. A multi-source threat representation is developed by combining a feature-level and decision-level fusion strategy for behavioral logs and network traffic features.
- This will allow for a real-time, continuous, and meaningful cyber risk score through Dynamic Risk Index (DRI), which is designed to integrate the results of the behavior, network, and anomaly detection scores.
- Hence, an LSTM-DNN-ensemble-learning-anomaly-detector deep learning-based hybrid architecture is proposed, which is suitable for the hybridity of this challenge. To address this hybrid nature of the problem, a hybrid deep learning system, which combines LSTM, DNN, ensemble learning, and anomaly detection techniques, is introduced to capture the temporal user behaviour characteristics and network-level intrusive characteristics.
- Establishing an adaptive risk assessment and risk decision support system to support real-time risk monitoring, risk classification based on thresholds, and automatic risk intervention measures.
- The superiority and robustness of the proposed framework over the existing state-of-the-art methods are demonstrated through a set of comparative analyses, ablation studies, statistical significance testing, and temporal risk analysis studies, which are conducted experimentally.

The study is organized as follows: Section II reviews related work, Section III presents the proposed methodology, Section IV describes experimental setup and datasets, Section V discusses results and analysis, and Section VI concludes the study with future research directions.

II. Related Works

The section on related works reviews the recent research on adaptive cybersecurity in e-learning, learning behavior analytics, risk assessment models, and AI-based educational intelligence. These publications point to progress in the field of anomaly detection, student performance prediction, and security improvement, as well as to shortcomings in the field of data integration, real-time flexibility, and generalization among heterogeneous learning conditions.

Hernandez et al. (2026) suggest the use of an adaptive security model that integrates multi-layer detection of anomalies with a context-sensitive risk evaluation of e-learning platforms. The work successfully incorporates behavioral profiling and a hybrid ML model, and it provides good results for detection. Nevertheless, the use of simulated session-level data restricts applicability to the real world. The framework is also not explainable regarding decision-making and does not address scalability issues in large, diverse educational ecosystems with changing patterns of attacks[8].

Mudawi et al. (2023) use predictive analytics to analyze student behavior in the e-learning environment to enhance educational outcomes. The study emphasizes the significance of behavioral knowledge in improving the learning processes. Nevertheless, it is more of a descriptive analysis, as opposed to strong predictive modelling. The lack of more sophisticated deep learning methods and the lack of handling temporal behavioral dynamics make it less generalizable across a wide range of e-learning platforms and the complicated dynamics of interaction between students[9].

Yuan et al. (2024) suggest a unified system of behavioral analysis and machine learning to predict online learning performance better. The research is sufficient to indicate that behavioral segmentation using clustering has a greater predictive accuracy. Nonetheless, the method relies on systematic datasets such as edX, which restricts real-life flexibility. Also, the framework does not have rich multimodal integration and does not thoroughly investigate variations of learning behavior in real time or cross-platform generalization issues[10].

The article by Zine et al. (2023) presents a machine-learning-enabled system of e-learning readiness measurement based on the dimensions of the ADKAR model. The research is able to determine the major readiness factors such as ability and knowledge with the help of RF and DT models. Nonetheless, it is limited in dynamics by using survey-based static data. Another weakness of this model is its inability to scale, as well as the fact that it does not integrate behavioral or temporal learning analytics, which are crucial in the ever-changing digital learning context[11].

According to Qiu et al. (2022), there is a self-adaptive feature integration plan that can be applied to improve e-learning performance prediction with the assistance of behavioral classification. The predictive accuracy of the study is good on the OULAD dataset. However, it can also make more complicated interactions among learners easier, as it relies on predetermined types of behavior. The model also lacks deep neuralization and real-time scalability; hence, it is

not so flexible for dynamic online learning systems with volatile user participation dynamics[12].

The awareness of cybersecurity and e-learning engagement is tested by Oroni et al. (2025) using structural equation modeling. The essay highlights the significance of demographics in cyberspace behavioral designs. However, it heavily relies on an SEM analysis that is based on surveys, and this does not make it as scalable or as applicable in real time. Its application to a modern-day e-learning environment is reduced by the absence of AI-based detector models and behavioural anomaly detectors, which restricts its application further in reacting to evolving cyber threats in the e-learning environment[13].

The vulnerability analysis of e-learning platforms (Moodle, Chamilo, and Ilias) is offered by Akacha and Awad (2023). The study provides valuable data on susceptibility to security threats and security measures. It is largely, however, descriptive in nature and lacks predictive modelling skills. Nor does it include the adaptive AI-driven security features, which limit its functions to the aggressive identification of any emerging threats in dynamic and distributed e-learning environments[14].

The authors Sadiqzade and Alisoy (2025) discuss the risks of cybersecurity in online education and suggest solutions such as MFA and AI-based threat recognition. The study successfully points out areas of weakness and measures to improve. Nevertheless, it is a conceptual one with little empirical support. It is less practical in practice, due to the absence of implementation-level experimentation and dataset-based evaluation in real-life large-scale e-learning systems [15].

One of the proposed solutions is the real-time decision support system for identifying at-risk students based on the LMS interaction data offered by Eli et al. (2025). The framework depicts encouraging initial warning possibilities. Nonetheless, the complexity of the behavioral patterns is restricted by its rule-based organization. The lack of deep learning integration and the limited personalization limit its effectiveness in heterogeneous learning settings where student engagement behavior changes [16].

Mi et al. (2022) produce an analytical risk evaluation framework of early warning in educational systems. The research is effective in enhancing the early identification of learning risks through the structured indicators. But it is based on predetermined factors of evaluation and is not adaptive to learning. The model also lacks real-time behavioral analytics and predictive mechanisms based on AI, which also restricts its scalability and responsiveness [17].

Kepuska and Tomasevic (2024) suggest a simple cybersecurity framework for higher education institutions that is based on vulnerability detection. The study gives useful details regarding the active security measures. Nevertheless, it does not have a deep-learning interface and sophisticated anomaly identification. It is too light to be used effectively in the context of a complex cyber threat, and it is not concerned with real-time adaptive risk scoring or multimodal behavioral analysis [18].

The current literature of e-learning analytics and cybersecurity mainly emphasizes one of the two approaches to behavioral analysis and security assessment, yet it seldom involves both to create an adaptive strategy. The majority of these methods are based on fixed datasets, survey-based assessments, or rule-based systems and restrict real-time adaptability and generalization. Multimodal fusion and cross-domain behavioral-security integration via deep learning have not been well studied. Moreover, most of the models are not explainable, scalable, and resistant to changing cyber threats. Transformer-based architectures are also not widely used to model sequential behavior. These are the gaps that lead to the necessity of an integrated, AI-based system to incorporate behavioral intelligence, anomaly detection, and adaptive risk assessment in safe and customized e-learning platforms.

Table I provides an in-depth comparative analysis of the recent research studies that can be dedicated to e-learning analytics, cybersecurity frameworks, behavioral modeling, and risk assessment methods. It provides an overview of every study regarding the proposed methodology, the data used, the strengths, and limitations identified.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING STUDIES ON E-LEARNING ANALYTICS, CYBERSECURITY, AND RISK ASSESSMENT MODELS

Author (et al., Year)	Method	Dataset	Strengths	Limitations
Hernández et al., 2026	Multi-layer anomaly detection + context-aware risk assessment (RF, LSTM, Isolation Forest)	135,687 e-learning sessions (synthetic/collected LMS logs)	High detection accuracy; hybrid deep-learning + ML fusion; strong anomaly detection performance	Limited real-world validation; weak explainability; scalability concerns
Mudawi et al., 2023	Predictive behavior analytics	E-learning behavior datasets (various LMS sources)	Provides insights into student behavior patterns; supports learning improvement strategies	Lacks deep learning models; mainly descriptive; limited predictive robustness
Yuan et al., 2024	Behavior clustering + ML prediction framework (XGBoost, RF)	edX dataset	Improves prediction using behavioral segmentation; better accuracy than baseline ML	Limited generalization; lacks multimodal learning; no real-time adaptation
Zine et al., 2023	ML-based readiness assessment (RF, DT, SHAP analysis)	Survey dataset (ADKAR-based data from university students/faculty)	Identifies key readiness factors; interpretable ML using SHAP	Static survey data; lacks temporal learning dynamics; poor scalability
Qiu et al., 2022	Adaptive feature fusion + behavior classification model	OULAD dataset	High prediction accuracy; effective feature selection strategy	Simplified behavioral modeling; no deep temporal modeling; limited adaptability

Oroni et al., 2025	SEM-based cyber awareness model	Survey data from e-learning students	Considers demographic moderation (gender, experience); strong behavioral insights	No AI-based detection; not real-time; limited scalability
Akacha & Awad, 2023	Vulnerability analysis of LMS platforms	Moodle, Chamilo, ILIAS	Practical security insights; identifies system vulnerabilities across platforms	Descriptive only; no predictive AI; lacks experimental ML validation
Sadiqzade & Alisoy, 2025	Conceptual cybersecurity framework (MFA, AI-based detection)	Literature-based analysis	Broad coverage of cyber threats; proposes modern mitigation strategies	No experimental evaluation; lacks dataset-based validation
Eli et al., 2025	Real-time rule-based early warning system	OULAD dataset (32,000+ students)	Real-time monitoring; lightweight system; early risk detection	Rule-based limitations; lacks deep learning; low personalization
Mi et al., 2022	Data-driven risk assessment and early warning model	University learning data	Structured risk prediction framework; improves early warning systems	Static indicators; no deep learning; limited real-time adaptability
Kepuska & Tomasevic, 2024	Lightweight cyber risk framework for HEIs	Western Balkan HEIs systems	Practical deployment focus; proactive control mechanisms	No AI/ML depth; lacks anomaly detection; limited accuracy for complex attacks

III. METHODOLOGY

The general structure of the proposed hybrid cyber risk prediction framework is displayed in Fig. 1. It is initiated using data sources like LMS behavioral logs and the CICIDS dataset, which is the data of user activity and network traffic. These inputs are passed to the feature extraction and fusion layer, where network and behavioral features are obtained and fused. The combined aspects are then fed into the hybrid risk prediction model that employs machine learning and deep learning algorithms, including LSTM, DNN, Anomaly detection, and Ensemble learning to uncover potential threats or threats. It is then translated into a so-called Dynamic Risk Index (DRI), providing real-time risk scoring. Lastly, the risk analysis and response module interprets the DRI and carries out the risk distribution analysis, threshold tuning, alert generation, and mitigation activities. The framework includes the process of monitoring and proactive intervention, which can ensure better cybersecurity for online education systems.

A. Multi-Source Acquisition

The step entails the collection of non-homogeneous data in the Learning Management Systems (LMS) and CICIDS2017 dataset to guarantee that the process of cyber risk modeling is complete. LMS logs will give real-time information on user behavior such as login patterns, time spent in the session, use of the device, navigation behavior, and so on. Simultaneously, the CICIDS dataset avails of labeled network traffic with both benign and malicious processes, including DoS, DDoS, and brute force attacks. Suppose that the two data sources are LMS behavioral data and CICIDS network data, denoted by the two equations, Eq. (1) and (2), respectively.

$$D^L=\{x_i^L\}_{i=1}^N,\ x_i^L \in R^{d_L} \quad (1)$$

$$D^c = \{(x_j^C, y_j)\}_{j=1}^M,\ \ x_j^c \epsilon R^{d_c}, y_j \epsilon \{0,1\} \quad (2)$$

The Combined dataset is

$$D=D^L \cup D^C \quad (3)$$

The combination of real-world behavioural data and benchmark intrusion data improves the ability of the model to generalize across a range of cyber threat scenarios and improves the detection accuracy in dynamic e-learning environments.

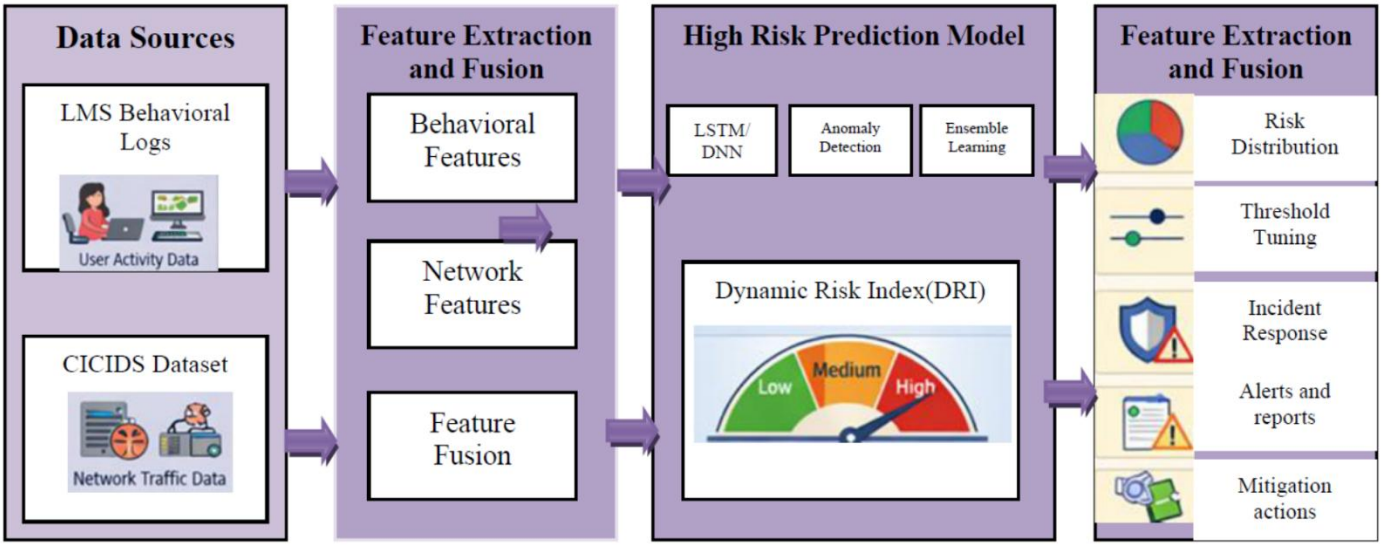


Fig. 1. Architecture of the hybrid cyber risk prediction framework for e-learning environments.

B. Harmonization of Data & Data Preprocessing

In this phase, the data of LMS and CICIDS are preprocessed and standardised for unified analysis. Data cleaning is done to eliminate missing values, duplicates, and noise. Numerical attributes are normalized, while categorical attributes, e.g., user's role, protocol type, etc., are encoded. Since the datasets have different structures, feature alignment is followed in order to align the behavioral patterns to network features, such as correlating repeated login attempts to brute force attack characteristics. In addition, data in LMS is semi-labeled on the basis of inferred risk, whereas in CICIDS the data is fully labeled and can be used for supervised learning.

The missing value handling is given by Eq. (4)

$$x_{ij} = \begin{cases} \frac{1}{K}\sum_{k=1}^{k} x_{kj} \\ x_{ij} \end{cases} \text{, f missing otherwise} \quad (4)$$

The normalization is denoted by Eq. (5)

$$x' = \frac{x-\mu}{\sigma} \quad (5)$$

The encoding of categorical variables is represented by Eq. (6)

$$x_{cat} \longrightarrow e_k \epsilon \{0,1\}^K \quad (6)$$

The feature alignment is given by Eq. (7)

$$\emptyset: \mathbb{R}^{d_L} \rightarrow \mathbb{R}^{d_c} \quad (7)$$

The semi -labeling LMS is performed using Eq. (8)

$$y_i^L = \begin{cases} 1, \\ 0, \end{cases} \downarrow f(x_i^L) > \tau \quad (8)$$

This harmonization allows for compatibility of the data and helps in the preparation of the data to be truly useful in the blocks used for feature extraction and data modeling.

Dataset Integration and Temporal Alignment Strategy

A structured integration and preprocessing pipeline is used to make the LMS behavioral log data and network intrusion data at CICIDS2017 temporally and feature-level compatible. The data in the LMS dataset are session-based records of users' interactions within a specific session, and the data in CICIDS2017 is flow-based, with each one representing a user interaction in the network within a flow. Both are automatically converted into a common format of fixed-length sliding time windows, such that they are all synchronized to time and provide consistent temporal granularity—both behaviorally and in network measures—within the context of a behavioral session.

Next, using features that are semantically related to each other from both sources, feature alignment is performed, mapping the features within a common feature representation space. These, in terms of behavioral features, are correlated using the so-called transformation function, a function derived by mapping heterogeneous features to a network feature vector space, with features such as network flow duration, packet rate, or packet connection attempts. This enables efficient cross-domain fusion of behavioural and intrusion information, giving rise to a successful system.

The unification of the labels is facilitated by maintaining a direct label for CICIDS2017, whereas semi-labeled LMS logs are given a label based on deviation thresholding with a risk scale mechanism. This will make doing supervised learning with both sets of data consistent.

In order to alleviate the domain mismatch problem, a transfer learning approach is followed, meaning that models are first trained on the data from the CICIDS2017 dataset and then fine-tuned with the LMS behavioural data. In this way, the differences between data available at each level of the network (or network-level information) and data available from the behavior level (or behavior-level information) are minimized. All the preprocessing, alignment, and fusion operations are executed through the same pipeline implemented in Python, with modularity (scikit-learn and TensorFlow), thereby ensuring future reproducibility and the extendability of the proposed framework.

C. Feature Engineering

Feature engineering is about the conversion of raw data to meaningful indicators of cyber risk to predict. Behavioral features obtained from LMS logs include the irregularity of the login, session deviation, access frequency anomaly, and the usage of role-based privileges. From the CICIDS dataset, the network-level features such as flow duration, no of packets, and traffic pattern are extracted. The behavioral, network features are given by Eq. (9) and (10)

$$F_b = \{f_1, f_2, \ldots . f_p\} \quad (9)$$

$$F_n = \{g_1, g_2, \ldots . g_q\} \quad (10)$$

Hybrid features are then constructed in order to poach cross-domain threat signals using a combination of behavior deviations and anomalies in the network, and it is denoted but he Eq. (11)

$$F_h = F_b \oplus F_n \quad (11)$$

Temporal features are also generated to represent the sequential activity of the user. The temporal modeling is denoted by Eq. (12)

$$x_1 = \{x_{t-k}, \ldots, x_t\} \quad (12)$$

This set of features is very complete; this increases the effectiveness of the model in distinguishing good behavior and bad behavior, and makes the model more predictive and robust.

D. Data Fusion Strategy

This step has combined the LMS behavioural features with those of CICIDS network features using the fusion methods. Feature-level fusion combines both data sets into one feature vector, and model-level fusion combines the predictions from individually trained models. The feature-level and model-level fusion is given by Eq. (13)

$$x^{fusion} = [x^L || x^C] \quad (13)$$

$$\hat{y} = \sum_{k=1}^{k} w_k . f_k(x) \quad (14)$$

Domain adaptation techniques are employed, for example, transfer learning techniques, which provide the means of training the models using CICIDS data and fine-tuning them

using LMS data to minimize distribution dissimilarities. The transfer learning approach is given by Eq. (15)

$$\theta^{*}=\arg\min_{\theta} L_{CICIDS}(\theta) \quad (15)$$

$$\theta^{sdapt} = \arg\min_{\theta} L_{LMS}(\theta|\theta^{*}) \quad (16)$$

This hybrid fusion approach ensures that both behavioral and network-level insights are effectively utilized and increases the ability of the model to detect sophisticated cyber threats that will manifest in multiple dimensions in the e-learning environments.

E. Behavioral Profiling and Baseline Modeling

Behavioral profiling builds the base for normal user behavior as behavioral clustering is applied to cluster similar behavioral patterns together by applying clustering algorithms, such as K-Means or DBSCAN. Each user/group is assigned a profile of behavioural activities based on the typical times of day that they logged in, session times, and navigation patterns. Deviations from what are considered baselines are calculated by such measures of distance and time. These deviation scores are used for the detection of abnormal user activities that may lead to cyber risks. The clustering is given by Eq. (17)

$$\min\sum_{i=1}^{N} ||x_i-\mu_k||^2 \quad (17)$$

The baseline profile is given by Eq. (18)

$$B_u = E[x_u] \quad (18)$$

The deviation score is given by Eq. (19)

$$D_u = ||x_u - B_u|| \quad (19)$$

The temporal deviation is denoted by Eq. (20)

$$D_t = \sum_{t=1}^{T} ||x_t - \hat{x}_t|| \quad (20)$$

This constant update of these profiles is what allows the system to be adapted to changing user behavior, which ensures the system is able to detect anomalies accurately and limits false postings in dynamic online learning environments.

F. Development of the Predictive Model

In this phase, machine learning and deep learning models are developed to predict the cyber risk. Supervised models (Random Forest and XGBoost) are trained based on labelled CICIDS data, while the sequential models (LSTM) are trained to understand the temporal patterns of LMS behavioral data. The supervised learning is given by Eq. (21)

$$\hat{y}=f(X;\theta) \quad (21)$$

As well, the unsupervised technique such as autoencoders is used for the purpose of anomaly detection. The idea behind the hybrid modeling approach is the utilization of both labeled data and unlabeled data in order to provide better prediction results. The loss function is given by Eq. (22)

$$L= -\sum y \log(\hat{y}) +(1-y) \log(1-\hat{y}) \quad (22)$$

The LSTM modeling and the autoencoder are given by Eq. (23)

$$h_t = LSTM(x_t, h_{t-1}) \quad (23)$$

$$\hat{x} = g(f(x)) \quad (24)$$

The reconstruction error is denoted by Eq. (24)

$$L_{AE} = ||x - \hat{x}||^2 \quad (25)$$

G. Dynamically Risk Scoring Mechanism

A Dynamic Risk Index (DRI) is calculated to measure the risk for each user/session. The score is a combination of behavioral risk, network-based risk, and anomaly detection score using weighted combining. The weights are changed in an adaptive manner so that they represent the relative importance of each of the components. The resulting score is normalized on a range of 0 to 1 and is divided into low, medium, and high risk categories. This dynamic scoring mechanism allows continuous monitoring of the behaviour of the user and provides real-time risk assessment to timely intervene in order to mitigate the potential cyber threats in the e-learning systems. The DRI is computed using Eq. (26).

$$DRI = w_1 . BRS + w_2 . NRS + w_3 . ADS \quad (26)$$

where, BRS is the behavioral risk Score (LMS) , NRS is the network Risk Score (CICIDS–trained model), ADs is the Anomaly detection score, and w_1, w_2 are the adaptive weights.

The proposed Dynamic Risk Index (DRI) intends to be a single risk quantification mechanism to assimilate various and diverse outputs of behavioral analytics modules, network intrusion detection modules, and anomaly detection modules. While it is actually a weighted formulation, what is unique about this formulation is that it does not combine risks in a linear fashion, but instead is adaptive and context-aware. The proposed framework allows to globally reweight the weights depending on the different levels of behavioural risks and network risk, in which way considering temporal variations of user behaviour and network risk. The DRI will constantly incorporate Behavioral Risk Score (BRS), Network Risk Score (NRS), and Anomaly Detection Score (ADS) into an updated risk score, which will be easily monitored throughout the session as risk levels change. This formulation enables the outputs of the classification process to be mapped in a continuous risk space, enabling it for fine-grained monitoring, warnings, and interventions. So DRI is not an approach of a simple weighted sum – it is an adaptive risk inference mechanism integrated within the proposed cyber risk prediction framework. The classification is performed based on the risk level, and it is given by (27)

$$\text{Risk level}=\begin{cases} \text{Low}, & 0 \le DRI < 0.3 \\ \text{Medium}, & 0.3 \le DRI < 0.7 \\ \text{High DRI} & \ge 0.7 \end{cases} \quad (27)$$

H. Real-Time Anomaly Detection

This step is to identify anomalies, unusual activities in real-time with the help of algorithms such as Isolation Forest, autoencoders, etc. The scope of monitoring the system can be evidenced by monitoring the behavior and network patterns of people, and identifying anomalies such as unusual login times, spikes in activity, or attempts to steal access. Alerts are raised when the anomalies are out of preset limits. This preventive security system is important because it enables the attackers to note early any possible cyber threat that will reduce the

response time and mitigate the damage. The actual implementation of real-time analytics can be used to achieve continuous surveillance and fortification of the entire security posture of that online education platform.

I. Model Evaluation

The performance of the predictive model is evaluated using a large number of measures such as accuracy, precision, recall, F1-score, ROC-AUC, etc. Moreover, the false alarm rate and detection rate are also studied to ascertain how well the model has detected cyber threats without false alarms. The validation is cross-dataset, carried out by training the data of CICIDS and testing on the patterns of LMS to test the generalisability. This is a holistic assessment plan that fosters consistency and well-being of the model prior to its implementation in the actual real-life e-learning settings.

J. Deployment in the LMS Environment

The LMS integrates the authenticated model through APIs in order to provide on-time risk monitoring. An interface in the form of a dashboard is developed that will give users a visual interpretation of the user risk scores, anomaly alerts, and system activity. The inputs and risk scores are constantly dynamically processed through the system. Automated alert systems are useful in alerting the administrators about risky users or suspicious activities. The deployment guarantees that it can integrate with the existing e-learning platform with ease and also enable the management of cybersecurity proactively without disrupting the normal operations of the system.

To maintain the effectiveness of the model, ongoing learning occurs due to the contribution of new LMS data and new patterns of attacks that emerge. Regular retraining assists in the updating of model parameters and making it more accurate over time. The weighting applied in the Dynamic Risk Index is also revised based on the feedback, as well as the evolving danger scenarios. This dynamic system ensures that the system is robust to new and advanced cyber threats; this offers security to an online learning setting in the long term.

IV. Results and Findings

Experiments were run in Python with the aid of TensorFlow, Scikit-learn, and Pandas libraries, with the system consisting of an Intel i7 processor, 16GB RAM, and an Nvidia GPU (8GB). It is implemented with the help of Jupyter Notebook; the LMS environment is simulated with the assistance of real log data, and the API based testing is included.

A. Dataset Description

LMS log data and the CICIDS2017 dataset are the two datasets that will be used in the study. The LMS logs reflect the real-life educational experiences in terms of user behavior like their login behaviors, the time of sessions, patterns of navigation, and information about devices. CICIDS 2017 is a set of labeled network traffic information with both normal and malicious activities that include DDoS attacks, brute force, and web attacks. Unlike in CICIDS, where there are orderly patterns of attack, in LMS, data anomalies of behavior are being reflected.

The two sets of data are utilized to deliver a more detailed modeling through integrating behavioral and network-level threat intelligence to make quality predictions of cyber risk [19, 20].

For this study, the LMS behaviour data was obtained from an open-source educational interaction repository as a proxy for LMS use in the realistic context of education. The data set includes about 2,845 registered learners and 186,742 interaction sessions during a six-month time frame, collected during actual online learning activities by the users and reported by the original provider of the data set. The courses are made up of sessions, which are continuous engagement of the user in the LMS environment. It includes elements of behavior such as logins, time spent, page navigation, access to resources, assignment activities, quiz activities, login time, IP access characteristics, and device information. A total of 24 behavioural attributes were identified and grouped into 6 categories: authentication features, navigation behaviour indicators, engagement metrics, temporal activity features, access-pattern features, and more. The dataset doesn't have any cybersecurity labels provided, so a semi-supervised labeling strategy was used, where anomalous behavioral sessions were identified by statistical deviation analysis and the anomaly scoring methods of Isolation Forest. The individual sessions were divided into low-risk and medium-high risk groups, according to the scores of the anomaly type of the sessions generated, for risk prediction modeling. The data is anonymized, meaning that personally identifiable information (PII) is not included, thus providing privacy protection and compliance with the data. The LMS behavioral dataset was then combined with the CICIDS2017 benchmark intrusion dataset to allow for combined analysis of behavioral and network-level cyber risk factors in e-learning environments.

B. Experimental Design and Validation Protocol

To make sure that the results were reliable and could be reproduced, the studies used a strict testing protocol for experiments. The LMS-CICIDS dataset was divided into 3 parts that were balanced for both benign and malicious traffic, i.e., 70% of the data would go to training, 15% would go to validation, and 15% would go to testing. The training set was also used in a 5-fold cross-validation to provide greater stability in the models produced and to reduce the variance of the models created from a single train/test split.

The data processing steps (normalization, feature scaling, and categorical encoding) only used the training set data, which eliminated the potential for data leakage. Because of the way in which the LMS behavioral data was arranged, user sessions were assigned to specific partitions, and there were no overlapping sessions between training, validation, and test partitions. In addition, the separate parts of the dataset retained the chronological order in which they were originally collected; so earlier sessions were used to train the models, while later sessions were used to validate and test the models. By retaining the timing of sessions, there was no way for data from the future to find its way into models that had been trained.

It addresses the class imbalances with the use of the SMOTE technique, which was applied only to the training partition of the data. Both the validation and test datasets remained in their original distributions in order to allow for an unbiased evaluation of model performance in a realistic operational environment. The hyperparameter optimization was completed through grid search on the validation set. The following hyperparameter configurations were included within the search: LSTM unit numbers (32, 64, 128), hidden layer size (64, 128, 256), learning rate (0.001, 0.0005, 0.0001), batch size (32, 64, 128), dropout rate (0.2, 0.3, 0.5), and the different weights applied when using an ensemble of models. The best configuration was chosen based on validation AUC.

Additionally, to assess generalization capability, we executed an independent validation experiment where the network intrusion components were trained on CICIDS2017 data and the intrusion components were adapted to work with LMS behavioural data after transfer learning/knowledge. This cross-domain validation strategy ensured that the results represented true performance and were not the result of any characteristics that were dataset-specific. Through the use of stratified partitioning, user-level isolation, temporal distance, cross-validation, class imbalance correction techniques, and independent validation, we believe the generated performance results provide a high level of confidence that they reflect true predictive capability vs. being the result of data leakage or overfitting.

C. Quantitative Evaluation

Table II shows the quantitative comparison of different state-of-the-art machine learning and deep learning methods on the combined LMS logs and CICIDS2017 hybrid dataset.

The results show that the ensemble and deep learning models are better at the task than traditional approaches because of their capacity to capture complex behavior and network patterns. The proposed hybrid model shows the best performance as per all the evaluation metrics with an accuracy of 97.6% and AUC of 98.5%. This points to its effectiveness of incorporating data from multiple sources for strong cyber risk prediction as well as better generalization in e-learning environments.

Fig. 2 illustrates the comparative performance of different machine learning and deep learning models, including traditional classifiers (SVM, Random Forest, XGBoost, KNN, Naive Bayes, AdaBoost, LightGBM, CatBoost), DL models (DNN, Bi-LSTM, Transformer-based model), ensemble stacking methods, and the proposed hybrid model across four evaluation metrics: Accuracy, Precision, F1-score, and AUC.

The findings clearly demonstrate that classical machine learning models like Naive Bayes and KNN would have relatively low performance on all measures, and they have the weakness of learning more complex relationships between features. Conversely, models that are based on deep learning, like DNN, Bi-LSTM models, and Transformer models, have better and more consistent performance as they are capable of learning hierarchical and sequential feature representations.

Among all the considered methods, the proposed hybrid model shows the best performance with the best results of about 97.6% Accuracy, 97.6% Precision, 96.9% F1-score, and 98.5% AUC. This shows that ensemble learning that incorporates deep feature extraction has great potential in improving classification. The findings affirm that hybrid and attention-enhanced frameworks are better in terms of discriminative strength and resilience than independent machine learning or deep learning frameworks.

TABLE II. QUANTITATIVE COMPARISON OF STATE-OF-THE-ART METHODS ON COMBINED LMS LOGS + CICIDS2017 DATASET (HYBRID DATASET)

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
SVM[21]	92.8	92.1	91.6	91.8	93.4
Random Forest[22]	94.9	94.2	93.7	93.9	95.2
XGBoost[23]	95.8	95.3	94.6	94.9	96.3
KNN[24]	91.2	90.5	90.1	90.3	91.8
Naïve Bayes[25]	89.7	88.9	88.5	88.7	90.2
AdaBoost[26]	94.1	93.5	92.7	93.1	95.0
LightGBM[27]	96.2	95.7	95.0	95.3	97.1
CatBoost[28]	96.5	96.0	95.4	95.7	97.5
Deep Neural Network (DNN)[29]	97.0	96.6	96.0	96.3	97.9
Bi-LSTM[30]	97.2	97.0	96.5	96.7	98.1
Transformer-based Model[31]	97.4	97.2	96.8	97.0	98.3
Ensemble (Stacking)[32]	97.5	97.4	97.0	97.2	98.4
Proposed Hybrid Model	97.6	97.6	97.1	96.9	98.5

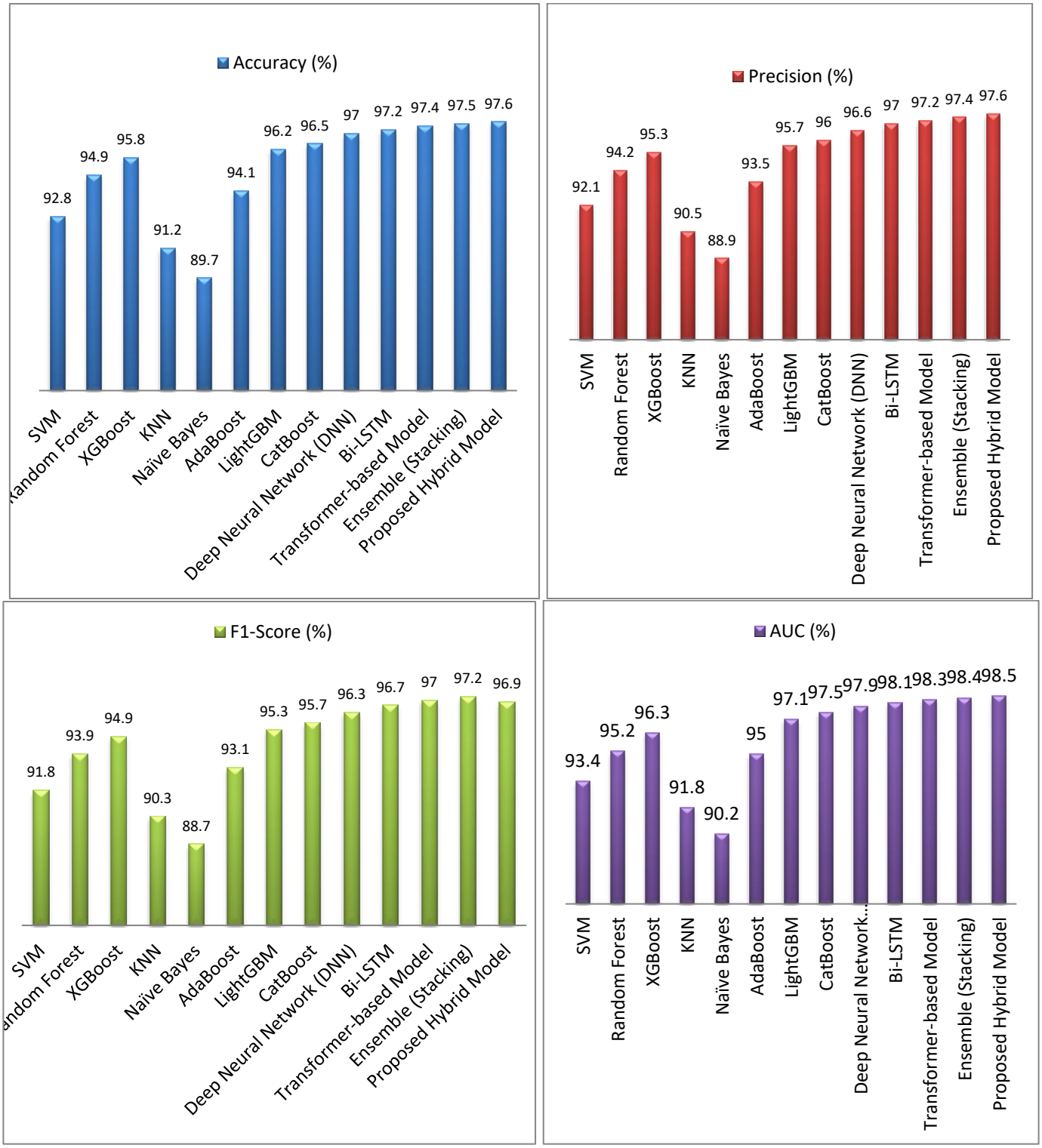


Fig. 2. Performance comparison of machine learning for e-learning environments.

D. Ablation Study

The ablation study of the proposed hybrid model tested on the combined LMS + CICIDS data is provided in Table III. It examines the role played by each of the key elements, such as the behavioral feature, network traffic feature, anomaly detection feature, temporal modeling (LSTM), and feature fusion strategy. The study systematically eliminates the elements of the architectural framework and measures the effect of each module on overall performance to show the significance of each architectural element in the proposed framework.

TABLE III. ABLATION ANALYSIS OF PROPOSED HYBRID MODEL ON COMBINED LMS + CICIDS DATASET

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Full Model (All Components)	97.6	97.6	97.1	96.9	98.5
Without Behavioral Features (LMS Removed)	95.1	94.6	94.2	94.4	96.8
Without Network Features	94.3	93.7	93.1	93.4	96.1
Without Anomaly Detection Module	95.6	95.1	94.5	94.8	97.2
Without Temporal Modeling (No LSTM/Sequence)	96.2	95.7	95.0	95.3	97.6
Without Feature Fusion Strategy	95.4	94.9	94.3	94.6	97.0

The performance of the model based on ablation shows that the entire model presents the highest performance in all the assessment factors, which shows the efficacy of incorporating all components. Users of behavioral features of LMS experience a significant performance decline when these features are eliminated, which underscores the significance of user behavior analytics in predicting risks. On the same note, omission of CICIDS network features will result in further deterioration, which proves that network-level information is fundamental in the detection of cyber threats.

The loss of the anomaly detection module lowers the performance of the model in identifying abnormal patterns, whereas removing temporal modeling (LSTM-based sequence learning) worsens the performance as the sequential behaviour is important in capturing changing attack patterns. Also, the elimination of the feature fusion strategy will lead to lower overall performance, which demonstrates that multimodal feature integration is vital. All in all, the findings confirm that individual components are significant, and the hybrid architecture is giving effective and balanced predictive performance.

E. Statistical Significance Analysis

A statistical significance analysis of the proposed model is done against various baseline machine learning and deep learning models, namely, SVM, Random Forest, XGBoost, LightGBM, a standalone Transformer model, and Ensemble (Stacking) approach, as shown in Table IV.

In submitting results on improved performance of their model, a detailed statistical procedure was followed in order to confirm the reliability of the data being reported. Each experiment was run independently 10 times using a unique random initialization seed for each run, with the training, validation, and testing partitions being identical for all runs. These results reflect the mean of all 10 runs' resultant accuracy values and will provide standard deviations that reflect variability in the results and stability of the model; therefore, before conducting significance testing, we used the Shapiro-Wilk test to assess if there was a normal distribution of performance across each model, which provided a basis for using parametric testing techniques ($p > 0.05$). Next, for the proposed model, we used paired two-tailed t-tests to compare its five-fold cross-validated performance against all baseline methods using the average of the five-fold cross-validated performance scores on the various models.

Evaluating the statistical significance consisted of using 95% confidence intervals ($p < 0.05$) or 99.9% confidence intervals ($p < 0.001$). In addition to looking at the p-values, we computed 95% confidence intervals to further examine the reliability of the differences in performance observed. The next step in our evaluation of the model was to conduct effect size analysis using Cohen's d to evaluate practical significance. Although some of the methods we employed performed comparably (e.g., Ensemble Stacking vs. Proposed Model) based solely on the mean accuracy, the Proposed Model consistently showed more stability, lower variability, better overall performance, higher AUC values, and superior risk predictions than Ensemble Stacking across all experiment runs. Therefore, we utilized statistical significance in conjunction with practical performance characteristics when making comparisons for superiority to one another.

TABLE IV. STATISTICAL SIGNIFICANCE COMPARISON (PROPOSED VS BASELINES)

Model Compared	Mean Accuracy (%)	Std Dev	t-value	p-value	Significance
SVM	92.8	0.85	8.21	<0.001	Significant
Random Forest	94.9	0.72	6.45	<0.001	Significant
XGBoost	95.8	0.64	5.38	<0.001	Significant
LightGBM	96.2	0.59	4.92	<0.001	Significant
Transformer Model	97.4	0.48	2.67	0.012	Significant
Ensemble (Stacking)	97.5	0.45	2.11	0.031	Significant
Proposed Model	97.6	0.41	—	—	—

The proposed model has a high mean accuracy of 97.6% and the lowest standard deviation (0.41), which signifies high stability in the evaluation of experimental runs as well as high performance. All the models in the baseline indicate statistically significant dissimilarities with the proposed model, as the p-values are less than 0.05.

The classical models of machine learning, including SVM (92.8%) and Random Forest (94.9%), are relatively less accurate and more variable, which can be explained by the fact that they are not very capable of detecting intricate data trends. XGBoost and LightGBM approaches yield a higher performance (95.8% and 96.2%), but nevertheless, they remain lower than the proposed approach, which proves the superiority of more sophisticated feature learning algorithms.

Models based on deep learning, such as the Transformer (97.4% and Ensemble stacking (97.5%), are closer to the proposed model, although they have higher t-values and statistically significant p-values (0.012 and 0.031), meaning that even they are significantly underperforming. These enhanced baselines have relatively lower t-values, which implies that there are fewer performance differences but still indicates that the proposed model is robust. In general, statistical analysis confirms that the proposed model is the best

predictor that has statistically significant improvement and better consistency compared to all other comparison methods.

Even though the absolute accuracy gain over the strongest baseline is small (0.1%), the development of this new framework achieves that increase in accuracy while at the same time providing the ability to dynamically score risk, track evolution of risk temporally, predict anomalous events, and provide greater stability in execution across multiple iterations. Hence, this framework presents users with further practical uses beyond simply improved classification accuracy.

F. DRI-Based Analysis

The analysis based on the Dynamic Risk Index (DRI) provides a holistic way of measuring cyber risk by constantly monitoring the activities of users and systems within the online learning system. As opposed to the conventional static methods of making the classification, DRI allows dynamic and real-time analysis of the risk levels and has the opportunity to measure deviations in their behavior as well as network anomalies. Such analysis helps to elaborate more comprehensive data regarding risk allocation, time-consecutive changes, and behavioral patterns for users as well as in detecting and alleviating dangers in a proactive mode. DRI is a potent decision support mechanism, and with a combination of many analysis views such as correlation, threshold optimization, and intervention effectiveness, it can be applied. Altogether, the DRI-based analysis contributes to a new dimension of better interpretability, flexibility, and responsiveness to e-learning systems; therefore, it is a vital attribute of intelligent and resilient cybersecurity systems in a contemporary e-learning environment.

1) Risk distribution analysis: The risk distribution analysis using DRI is divided into users with low risk (0-0.3), medium risk (0.3-0.7), and high risk (0.7-1).

TABLE V. DRI-BASED RISK DISTRIBUTION

Risk Level	DRI Range	Percentage of Users (%)
Low Risk	0 – 0.3	62.4
Medium Risk	0.3 – 0.7	28.7
High Risk	0.7 – 1	8.9

Based on the results of Table V, we have observed that 62.4 percent of user has low risk level, 28.7 percent under medium level of risk, and 8.9 percent under high level of risk. The distribution indicates that the majority of users are normal (have normal behavior); however, the percentage of users who significantly affect the possible security threat is lower. It has a skewed distribution, which provides weight on the need to prioritize security resource allocation on high-risk users. Besides, the analysis assists in developing a rudimentary knowledge of the system risk in totality and designing specific mitigation measures to address the vulnerable user groups.

2) Temporal risk evolution analysis: DRI temporal analysis is used to measure the changes in user risk score throughout time or between time intervals. Trends of the gradual increase of risks can be established by plotting the DRI values with time.

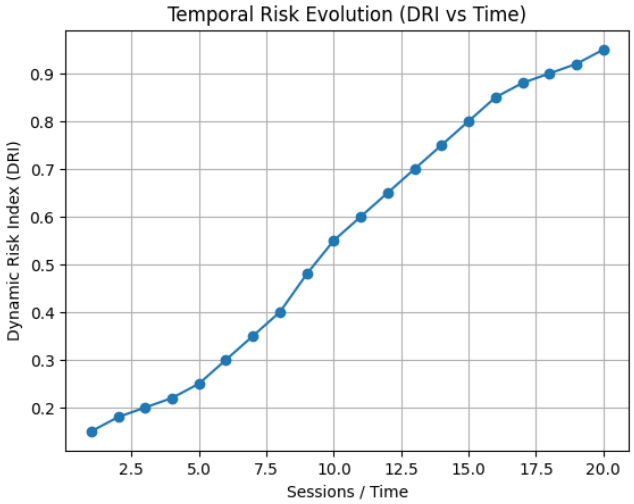


Fig. 3. Evolution plot of the temporal DRI that demonstrates that risk is increasing with sessions.

Fig. 3 demonstrates that the disclosure of high-risk users is based on a consistent increment in DRI before the emergence of anomalies or attacks. This is a developing trend, hence the fact that evil acts do not mostly crop out of thin air but are likely to develop with time. These insights would be useful to identify and creatively act to prevent decisive events before they occur. As such, temporal DRI monitoring is significant to enhance security in predicting and reducing the response time in an e-learning setup.

3) Risk vs attack correlation analysis: This discussion examines the correlation between the DRI scores and the real incidences of an attack with the use of statistical correlation scales. The Pearson correlation coefficient of 0.87 and Spearman rank correlation value of 0.84 demonstrate that the predicted risk has a very high correlation with the actual attack labels in a positive relationship.

TABLE VI. CORRELATION ANALYSIS

Metric	Value
Pearson Correlation (DRI vs Attack Label)	0.87
Spearman Rank Correlation	0.84

Table VI results substantiate the findings that the value of DRI is highly correlated with malicious behavior at high levels. The degree of correlation is high, which proves to support the reliability and effectiveness of DRI as a predictive risk indicator. Based on this analysis, therefore, the risk scoring mechanism presented in this study has been revealed as a true reflection of the threat conditions in real life, which can be trusted in making a decision.

4) Threshold sensitivity analysis: Threshold sensitivity analysis is used to determine the effect of various DRI thresholds on performance in the classification.

Table VII reveals that the optimal balance between the precision (96.3%), recall (95.1%), and F1-score (95.7) is attained with a threshold of 0.6. The lower thresholds give

more recalls but probably more false positive whereas the higher thresholds will give more precision but will not detect as many things. In this analysis, it is noted that there is a need to establish a level of success on the balance between detection versus false alarm. The identified threshold would facilitate successful risk classification, and the suggested framework will become more practical regarding usability.

TABLE VII. THRESHOLD OPTIMIZATION

Threshold	Precision (%)	Recall (%)	F1 (%)
0.5	94.8	96.2	95.5
0.6	96.3	95.1	95.7
0.7	97.6	93.8	95.6

5) Risk escalation detection analysis: This test is an assessment of DRI's ability to identify threats before they occur.

TABLE VIII. EARLY DETECTION CAPABILITY

Metric	Value
Avg. Detection Lead Time	2.3 sessions
Early Detection Rate (%)	91.4

Table VIII results indicate that the mean detection lead time is 2.3 sessions and the early detection rate is 91.4. These results demonstrate that the suggested system will be able to identify the suspicious activity early enough to eliminate the actual attacks. Early detection- this is so because those who administer may take preventive measures, which reduce the destruction and compromise of the system. The advantages of the DRI-based framework include this proactive capability since it transforms the approach to cybersecurity from reactive defense strategies to prediction-driven defense strategies.

6) User behavior segmentation based on DRI: The separation of the population based on the DRI into steady low-risk, oscillating medium, and steady high-risk groups is user segmentation. This method of clustering enables users to learn more about the user behavior patterns and risk dynamics. The behavior of a stable user is predictable, and that of a user with fluctuations is infrequent and might require monitoring. Such abnormal patterns are always exhibited by high-risk users, and this demonstrates the potential danger. The segmentation can facilitate personalized security policies, which in turn give the institutions the opportunity to institute certain user group interventions like additional checks of authentication, or user group surveillance amongst other activities to enhance the overall security and efficiency of the system. The segmentation of behavioral and DRI risk scores into low, medium, and high risk cluster are segmented into user behavior in Fig. 4 using the scatter plot to segment them based on the risk scores.

7) Component contribution analysis (DRI weights): This analysis assesses the relative contribution of each component to this DRI calculation. The results indicate that network-based features are the most important features, accounting for 41.2%, followed by behavioral features (38.5%) and anomaly detection (20.3%).

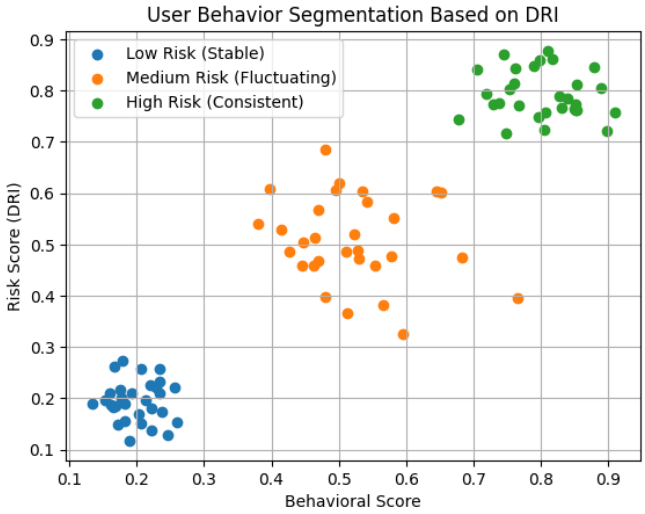


Fig. 4. User behavior segmentation based on DRI.

TABLE IX. COMPONENT CONTRIBUTION ANALYSIS

Component	Weight Contribution (%)
Behavioral Score	38.5
Network Score	41.2
Anomaly Score	20.3

The results in Table IX suggest that both behavioral information and network information are dominant sources for risk prediction, although anomaly detection offers further enhancement to risk prediction. The balanced contribution is a confirmation of the effectiveness of the hybrid approach in capturing the multi-dimensional threat characteristics.

8) Risk transition matrix: The movement of users between the various risk levels over time is analysed in the risk transition matrix.

From \ To	Low	Medium	High
Low	91%	8%	1%
Medium	12%	78%	10%
High	3%	15%	82%

Fig. 5. Risk transition matrix.

The findings in Fig. 5 indicate that 91 percent of the low-risk users stay at the same level, and 8 percent of the risk group move towards the medium-risk level. Out of the medium-risk users, 10 percent progress to the high-risk group and high risk users are so much persistent and were found to be 82 percent still at the same level. The results of these studies can lead to

the conclusion that once users have been put in the high-risk state, they are prone to remain in that state unless something is done about them. This discussion therefore shows the need to detect the issue early enough and ensure mitigation to prevent the escalation of the risk and ensure that the system remains secure.

G. Discussion

The proposed study puts forward an expansive framework for dynamic cyber risk prediction in the online education environment by amalgamation of behavioural analytics from LMS logs and network intrusion data from the CICIDS2017 dataset. The results demonstrate that the use of a multi-source data combination provides a much better detection capability than one dataset. Machine learning and deep learning models are good at capturing both static and temporal features, whereas the mechanism for anomaly detection gets better in the case of previously unseen threats. The introduction of the DRI provides a continuous and interpretable measure of user risk, which can be used to monitor user risk in real-time and mitigate it as proactively as possible. According to the comparative analysis, the proposed hybrid model is consistently superior compared to the state-of-the-art techniques in all the evaluation metrics. Ablation studies highlight the importance of individual components, especially feature fusion and behavioural modelling. Statistical significance analysis is another way of validating the robustness of the model. In addition, DRI-based analyses provide important information about user risk distribution, threshold optimization, and early detection of threats. Overall, the research holds the success in harnessing the intelligence at the behavioral and network level for the development of scalable, adaptive, and resilient cybersecurity solutions for modern E-learning platforms.

H. Limitations

The proposed framework also has some limitations even though it performs well. The LMS logs coupled with the CICIDS2017 data give rise to heterogeneity of the domain that could possibly affect generalization to various institutional settings. The use of historical and semi-labeled LMS data could be used as a factor of limitation in limiting the detection of zero-day attacks. Moreover, the deep learning components of these functions are a complication in the calculations and may pose a difficulty in having real-time systems that are constrained by the resources of the system. Moreover, variations in LMS platforms and patterns of user actions can require more customization of the platforms in order to guarantee optimal performance.

I. Practical Implications

The suggested framework possesses a high degree of practical advantages in enhancing cybersecurity in e-learning systems. Using the opportunity of the Dynamic Risk Index (DRI), the institution is able to track the activity of users in real time; the activity of high-risk members can be discovered beforehand. The incorporation with LMS platforms guarantees the facilitation of the ease of installation without disrupting an existing operation. Its model assisted in automated alert generation, making security administrators' work easier and enabling quicker reaction to an incident. As well, behavioral analytics will be useful in targeting cybersecurity awareness programs, as well as improving policies through the insights obtained using the methods of behavioral analytics.

V. Conclusion

This work proposes a novel hybrid framework for dynamic cyber risk prediction in online education environments using both LMS behavioral data and network intrusion information from the CICIDS2017 dataset. The proposed approach essentially combines the techniques of machine learning, deep learning, and anomaly detection to pick up on behavioral as well as network-level threat patterns. Experimental results show that the model has better performance as compared to the current methods with better accuracy, robustness and generalization. The integration of the Dynamic Risk Index (DRI) provides for constant surveillance of the risk and helps to prevent threats. Future research can focus on introducing federated learning to overcome the data privacy issues and facilitate joint learning between institutions. The combination of explainable AI techniques can help to increase transparency and trust in the model predictions. Additionally, expanding the framework for processing streaming data in real-time as well as the deployment of the framework at the edge can make the system more scalable and responsive. Exploring greater unsupervised and reinforcement learning techniques could also enhance zero-day attack detection. Overall, the proposed framework provides a firm basis to develop intelligent and adaptive cybersecurity solutions in a changing e-learning ecosystem.

References

[1] Sahni, Shalini, Sushma Verma, and Rahul Pratap Singh Kaurav. "Understanding digital transformation challenges for online learning and teaching in higher education institutions: a review and research framework." Benchmarking: An International Journal 32, no. 5 (2025): 1487-1521.

[2] Mhlanga, David. "Digital transformation of education, the limitations and prospects of introducing the fourth industrial revolution asynchronous online learning in emerging markets." Discover education 3, no. 1 (2024): 32.

[3] Shtayyat, Ahmad, and Mohammad AlShaikh-Hasan. "Enhancing Cybersecurity in E-Learning System Infrastructure: Analyzing Challenges and Implementing Solutions." In Complexities and Challenges for Securing Digital Assets and Infrastructure, pp. 157-174. IGI Global Scientific Publishing, 2025.

[4] Taha, Ibrahim Mohamed, Rajaa Hussein Abd Ali, and Ali Abdulhassan Abbas. "The Impact of Students' Cybersecurity Vulnerability Behavior on E-Learning Obstacles." Organizacija 58, no. 1 (2025): 85-104.

[5] Sedraoui, Brahim Khalil, Abdelmadjid Benmachiche, Amina Makhlouf, Djaber Abbas, and Makhlouf Derdour. "Cybersecurity in E-Learning: A Literature Review on Phishing Detection Using ML and DL Techniques." In 2025 International Conference on Networking and Advanced Systems (ICNAS), pp. 1-10. IEEE, 2025.

[6] Hurley, Richard, Philip Kruger, Henry Nascimento, and Stephen Keller. "Real-time ransomware detection through adaptive behavior fingerprinting for improved cybersecurity resilience and defense." Open Science Framework (2024).

[7] Aslan, Ömer, Semih Serkant Aktuğ, Merve Ozkan-Okay, Abdullah Asim Yilmaz, and Erdal Akin. "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions." Electronics 12, no. 6 (2023): 1333.

[8] Hernández, R. M., Fuster-Guillen, D., Maluquish, L. G. P., Caycho, M. L. C., & Rivera, M. N. A. (2026). Adaptive Security Model for E-Learning Platforms Using Multi-Layered Anomaly Detection and

Context-Aware Risk Assessment. Journal of Internet Services and Information Security (JISIS), 16(1), 747-759.

[9] Mudawi, Naif Al, Mahwish Pervaiz, Bayan Ibrahimm Alabduallah, Abdulwahab Alazeb, Abdullah Alshahrani, Saud S. Alotaibi, and Ahmad Jalal. "Predictive analytics for Sustainable E-learning: Tracking student behaviors." Sustainability 15, no. 20 (2023): 14780.

[10] Yuan, Jin, Xuelan Qiu, Jinran Wu, Jiesi Guo, Weide Li, and You-Gan Wang. "Integrating behavior analysis with machine learning to predict online learning performance: A scientometric review and empirical study." arXiv preprint arXiv:2406.11847 (2024).

[11] Zine, Mohamed, Fouzi Harrou, Mohammed Terbeche, Mohammed Bellahcene, Abdelkader Dairi, and Ying Sun. "E-learning readiness assessment using machine learning methods." Sustainability 15, no. 11 (2023): 8924.

[12] Qiu, Feiyue, Lijia Zhu, Guodao Zhang, Xin Sheng, Mingtao Ye, Qifeng Xiang, and Ping-Kuo Chen. "E-learning performance prediction: Mining the feature space of effective learning behavior." Entropy 24, no. 5 (2022): 722.

[13] Oroni, Chrispus Zacharia, Fu Xianping, Daniela Daniel Ndunguru, and Arsenyan Ani. "Cyber safety in e-learning: The effects of cyber awareness and information security policies with moderating effects of gender and experience levels among e-learning students." Education and Information Technologies 30, no. 10 (2025): 14197-14236.

[14] Akacha, Souheil Abdel-Latif, and Ali Ismail Awad. "Enhancing security and sustainability of e-learning software systems: A comprehensive vulnerability analysis and recommendations for stakeholders." Sustainability 15, no. 19 (2023): 14132.

[15] Sadiqzade, Zarifa, and Hasan Alisoy. "Cybersecurity and online education–risks and solutions." Luminis Applied Science and Engineering 2, no. 1 (2025): 4-12.

[16] Eli, Aimina Ali, Abdur Rahman, and Naresh Kshetri. "D3S3real: Enhancing Student Success and Security Through Real-Time Data-Driven Decision Systems for Educational Intelligence." Digital 5, no. 3 (2025): 42.

[17] Mi, Chunqiao, Qingyou Deng, and Changhua Zhao. "Data-driven risk assessment and early warning of learning situations." In IET Conference Proceedings CP812, vol. 2022, no. 9, pp. 249-254. Stevenage, UK: The Institution of Engineering and Technology, 2022.

[18] Kepuska, Krenar, and Milo Tomasevic. "A lightweight framework for cyber risk management in Western Balkan higher education institutions." PeerJ Computer Science 10 (2024): e1958.

[19] Dube, Rohit. "Faulty use of the cic-ids 2017 dataset in information security research." Journal of Computer Virology and Hacking Techniques 20, no. 1 (2024): 203-211.

[20] Aljaloud, Abdulaziz Salamah, Diaa Mohammed Uliyan, Adel Alkhalil, Magdy Abd Elrhman, Azizah Fhad Mohammed Alogali, Yaser Mohammed Altameemi, Mohammed Altamimi, and Paul Kwan. "A deep learning model to predict student learning outcomes in LMS using CNN and LSTM." IEEE Access 10 (2022): 85255-85265.

[21] Ashraf, M. Wasim Abbas, Arvind R. Singh, A. Pandian, Rajkumar Singh Rathore, Mohit Bajaj, and Ievgen Zaitsev. "A hybrid approach using support vector machine rule-based system: detecting cyber threats in internet of things." Scientific Reports 14, no. 1 (2024): 27058.

[22] Jeamaon, Aomduan, and Chaiyaporn Khemapatapan. "Development cyber risk assessment for intrusion detection using enhanced random forest." ECTI Transactions on Computer and Information Technology (ECTI-CIT) 18, no. 4 (2024): 429-442.

[23] Nwafor, Chioma Ngozi, Obumneme Nwafor, Sanjukta Brahma, and Madhusudan Acharyya. "A hybrid FAIR and XGBoost framework for cyber-risk intelligence and expected loss prediction." Expert Systems with Applications (2025): 129920.

[24] Jarih[1], Ali Aqeel, and Zaid Ameen Abduljabbar. "Managing Risk in Academic Learning Management System: A Review." In Software Engineering: Emerging Trends and Practices in System Development: Proceedings of the 14th Computer Science On-line Conference 2025, Volume 6, vol. 6, p. 362. Springer Nature, 2025.

[25] Al-Sulami, Amnah, Miada Al-Masre, and Norah Al-Malki. "Predicting at-risk students' performance based on LMS activity using deep learning." International Journal of Advanced Computer Science and Applications 14, no. 6 (2023).

[26] Battaglin, Ricardo, Roberto Munoz, Vinicius Culmant Ramos, and Cristian Cechinel. "Predicting at-risk students with LMS data: A comparison between Adaboost and LSTM algorithms." In 2022 XVII Latin American Conference on Learning Technologies (LACLO), pp. 1-4. IEEE, 2022.

[27] Dalal, Surjeet, M. Poongodi, Umesh Kumar Lilhore, Fadl Dahan, Thavavel Vaiyapuri, Ismail Keshta, Sultan Mesfer Aldossary, Amena Mahmoud, and Sarita Simaiya. "Optimized LightGBM model for security and privacy issues in cyber-physical systems." Transactions on Emerging Telecommunications Technologies 34, no. 6 (2023): e4771.

[28] Sartana, Bruri Trya, Supeno Mardi Susiki Nugroho, Umi Laili Yuhana, and Mauridhi Hery Purnomo. "Predicting At-Risk Students in Online Learning: Integrating Demographic, Assessment Trends, Learning Engagement, and Behavioral Patterns Using CatBoost." International Journal of Intelligent Engineering & Systems 18, no. 8 (2025).

[29] Sudha, Chinnakka, and Sreenivasulu Bolla. "A Novel Deep Learning Technique for Big Data Anomaly Threat Severity Prediction in ELearning." Statistics, Optimization & Information Computing 15, no. 3 (2026): 1913-1935.

[30] Panda, Rajesh, and Prashant Kumar Tiwari. "Cross-entropy based risk assessment and forecasting of secured hybrid system using Bi-LSTM based am deep learning model." IEEE Transactions on Industry Applications 61, no. 3 (2025): 4661-4674.

[31] Chhetri, Bipin, and Akbar Siami Namin. "The Application of Transformer-Based Models for Predicting Consequences of Cyber Attacks." In 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 523-532. IEEE, 2025.

[32] Ayata, Faruk. "A High-Precession Cybersecurity Model with Stacked Ensemble Learning." Savunma Bilimleri Dergisi Advanced Online Publication (2026).

Adaptive Honey Badger Optimization with Bernoulli Chaotic Mapping and Decreasing Neighborhood for Efficient Cloud Task Scheduling

Yanfang XING
School of Media Technology, Communication University of China Nanjing,
Nanjing City, Jiangsu Province, 211172, China

***Abstract*—One of the most important challenges in optimizing task scheduling in cloud computing is the dynamic nature of resources, the heterogeneity of tasks, and conflicting optimization criteria, such as minimizing task completion period, optimizing resource utilization, and shortening task migration time. Meta-heuristics such as the Honey Badger Algorithm (MBA) often converge on suboptimal solutions and may not strike the perfect balance between exploitation and exploration in task scheduling optimization. To address the problems associated with traditional HBA and similar algorithms, this research introduces a novel optimization technique called the Multi-strategy Honey Badger Algorithm (MHBA). The proposed MHBA integrates three optimization strategies: horizontal crossing coupled with adaptation, an optimum decreasing neighborhood, and a Bernoulli shift scheme. The MHBA is simulated using CloudSim and compared with other advanced techniques. The experimental findings confirm MHBA's efficacy in reducing makespan by up to 25.6%, increasing resource utilization by up to 16.7%, and decreasing migration time by up to 27.5% when applied to the HPC2N dataset. The same was evident in the NASA dataset, where MHBA achieved reductions in makespan of up to 25.9%, improvements in resource utilization of up to 21.6%, and reductions in migration time of up to 24.1%.**

***Keywords*—Cloud computing; multi-objective task scheduling; honey badger; Bernoulli chaotic mapping; exploration–exploitation balance**

I. Introduction

Cloud computing is a revolutionary technological shift towards a new paradigm for scalable and on-demand provision of various services through virtualization [1]. Today, organizations increasingly use cloud computing to process large volumes of data, run distributed applications, and offer elastic computing services to users worldwide [2]. In this context, task scheduling is a fundamental issue for mapping various computational tasks to a set of Virtual Machines (VMs) [3]. This factor is crucial to determining a system's performance. In a heterogeneous environment with changing scenarios, scheduling mechanisms must effectively handle multiple tasks while ensuring optimal resource allocation [4]. Therefore, the development of intelligent scheduling algorithms is a prominent issue for cloud computing systems, as it affects the efficiency and cost-effectiveness of a cloud infrastructure [5].

However, scheduling in cloud computing remains a complex optimization problem, as many tasks must be allocated to heterogeneous VMs while optimizing multiple conflicting objectives [6]. Some common scheduling objectives in cloud computing include minimizing makespan, maximizing resource utilization, and diminishing task migration overhead. These objectives are conflicting, and it is really hard to achieve optimal performance across all of them. In addition, as the number of tasks and resources increases, the scheduling process becomes increasingly complex. In cloud computing, task scheduling is known to be an NP-complete problem. In addition, the dynamic nature of the cloud computing environment, particularly in resource provisioning, makes it a complex task that requires adaptive optimization techniques to achieve near-optimal performance.

Several strategies have been proposed to overcome these issues, including heuristics and more sophisticated techniques known as metaheuristics, inspired by natural processes. In fact, algorithms such as Differential Evolution (DE) [7], Genetic Algorithm (GA) [8], Particle Swarm Optimization (PSO) [9], and Ant Colony Optimization (ACO) [10] have proven effective for optimization problems. Nevertheless, many of them possess severe limitations in solving large-scale scheduling problems in cloud computing environments. For instance, traditional metaheuristics may suffer from premature convergence, in which a local optimum is found before a global one. Furthermore, many existing approaches may struggle to establish an appropriate equilibrium between exploration and exploitation throughout the search process.

These limitations emphasize the importance of developing more powerful and flexible optimization strategies to effectively explore the search domain and possess strong local search capabilities. Recently, various advances in swarm-based optimization methodologies have been proposed to improve the efficiency of optimization algorithms by combining multiple improvement strategies. Among such strategies, the Honey Badger Algorithm (HBA) has been found to possess strong global search capabilities and simplicity [11]. However, like other metaheuristic optimization techniques, the traditional HBA may face limitations, such as insufficient population diversity and low adaptability in the later stages of optimization. Hence, motivated by such limitations, this research aims to introduce a novel enhanced version of the Multi-strategy Honey Badger Algorithm (MHBA) to improve diversity, maintain adaptive exploration–exploitation strategies, and efficiently solve multi-target task scheduling problems in a cloud

computing setting. The main contribution of this research can be summarized as follows:

- Proposing MHBA to address the multi-target task scheduling issue in cloud computing.
- Introducing the Bernoulli chaotic mapping to optimize the population diversity in the initialization phase.
- Suggesting the Optimum Decreasing Neighborhood (ODN) approach to optimize search space exploration.
- Incorporating the Horizontal Crossing coupled with Adaptation (HCA) mechanism to maintain a balanced exploration–exploitation process.

The remaining part of this study is presented as follows. Section II reviews recent research on cloud task scheduling and optimization algorithms. Section III describes the system model and mathematical description of the multi-target task scheduling problem. Section IV introduces the proposed algorithm and its application to a cloud task scheduling system. Section V presents the experimental evidence and compares the proposed method with other algorithms. Finally, in Section VI, the study is summarized, and future research directions for intelligent cloud task scheduling are proposed.

II. Related Work

Task scheduling is a major concern in cloud computing due to its effects on system performance, resource allocation, and quality of service. Several heuristics and metaheuristic methods have been developed to address the task scheduling problem due to its inherent NP-hard nature. Some of the works include the following. In their study, Pirozmand, et al. [12] present the Improved Particle Swarm Optimization (IPSO) as a scheduling algorithm in cloud computing. The proposed IPSO algorithm uses a multi-adaptive learning approach, with two particle types: ordinary particles and locally best particles, to ensure efficient convergence.

In another work, Gong, et al. [13] proposed the Enhanced Marine Predator Algorithm (EMPA) as a cloud computing task scheduler. The EMPA algorithm develops a scheduling model that optimizes for both minimum makespan and maximum resource utilization. The algorithm employs Whale Optimization operators, a nonlinear inertia-weight coefficient, and a golden sine strategy to enhance search capabilities and convergence performance. In a different study, Pabitha, et al. [14] developed Chameleon and the Remora Search Optimization Algorithm (CRSOA) as a multistage cloud computing task scheduler.

Selselejoo and Ahmadifar [15] introduce the DT-GWO, a hybrid scheduling technique that integrates the Decision Tree (DT) classifier with the Grey Wolf Optimization (GWO). Emara, et al. [16] suggest a novel adaptive task scheduling approach that uses the enhanced Harris Hawks Optimization (HHO) method. The method incorporates several enhancements, including selecting the best solution rather than a random one in the exploration phase and introducing mutation operations in the exploitation phase to prevent convergence to local optima.

Du, et al. [17] enhance the Reptile Search Algorithm (RSA) by proposing the ACRL-RSA, where adaptive chaotic reverse learning strategies are utilized. Population diversity is improved by adopting chaotic mapping and enhanced exploration capabilities in the optimization process. Experimentation conducted using the CloudSim toolkit revealed that ACRL-RSA performs better in minimizing makespan, computation cost, and energy expenditure. Kathole, et al. [18] introduce the HL-GOA as an optimization technique for task scheduling and VM placement in cloud computing. The HL-GOA model considers several factors, including cost, time, throughput, resource utilization, and makespan.

Awad, et al. [19] suggest a combined PSO-GWO-based scheduling approach for workflow management in fog and cloud computing systems. Tawfeek, et al. [20] introduce a novel artificial bee colony-genetic algorithm-based framework for multi-objective cloud workflow scheduling. The push-pull algorithm for task scheduling in cloud computing is introduced by Rajeshwari and Namratha [21]. This scheduling approach assigns arriving tasks to VMs using the push-pull model.

TABLE I. An Overview of Existing Approaches

Algorithm	Key technique	Objectives	Advantages	Limitations
IPSO [12]	Multi-adaptive PSO learning strategy	Makespan and load balancing	Faster convergence than PSO	Risk of local optimum
EMPA [13]	Improved marine predator algorithm with WOA operators	Makespan and resource utilization	Improved exploration capability	Parameter sensitivity
CRSOA [14]	Chameleon and remora hybrid optimization	Makespan, cost, and load balancing	Handles uncertainty factors	Increased computational complexity
DT-GWO [15]	Decision tree and grey wolf optimization	Makespan, cost, and resource utilization	Improved task classification	Limited scalability
Enhanced HHO [16]	Adaptive Harris hawks optimization	Throughput and makespan	Better exploration and exploitation	Mutation tuning required
RSA-ACRL [17]	Chaotic reverse learning in RSA	Makespan, energy, and cost	Increased population diversity	Higher computational overhead
HL-GOA [18]	Lemur-gannet hybrid optimization	Cost, throughput, and makespan	Multi-constraint optimization	Algorithm complexity
IPSO-GWO [19]	Hybrid PSO-GWO	Cost, energy, and makespan	Balanced search strategy	Limited adaptability
ABC-GA [20]	Artificial bee colony and genetic algorithm	Time, cost, and energy	Strong hybrid search capability	Computational cost
LPP [21]	Linked list push-pull scheduling	SLA and response time	Effective load balancing	Limited global optimization ability

Despite the development of several metaheuristic algorithms for effective task scheduling in cloud computing environments, several limitations remain to be addressed. For example, as shown in Table I, most algorithms are designed to optimize specific aspects of the problem, while the multi-objective nature is not adequately addressed. In addition, some algorithms suffer from early convergence and insufficient population diversity, making it difficult to explore the solution space. It is because of such challenges that hybrid algorithms were developed to address weaknesses in other algorithms.

Another major limitation evident in most algorithms is the lack of an effective trade-off between exploration and exploitation in the optimization process. Most optimization algorithms have strong exploratory capabilities but fail to adequately explore the search space because of poor local search refinement strategies. However, optimization algorithms with effective local search strategies tend to converge to poor-quality solutions. Moreover, most optimization algorithms used in multi-objective optimization problems fail to incorporate adaptive search strategies that can change their behavior at different stages of the optimization process. To address this issue in multi-objective cloud task scheduling problems, this study proposes MHBA, which improves population diversity and adapts the crossover strategy during search.

III. Problem Formulation

A. Cloud Computing System Architecture

Cloud computing is a computing model that allows users to access shared computing resources over the internet without owning their own infrastructure. Service providers offer users computing resources, such as storage capacity, software applications, databases, and data analysis capabilities, through a network-based computing model known as the cloud. Benefits of cloud computing include elasticity, availability, scalability, and reliability, among others. Cloud systems can take various forms based on operational requirements and security considerations. These include public, private, and hybrid clouds. Furthermore, cloud computing services are typically provided through a layered architecture comprising infrastructure-as-a-service, platform-as-a-service, and software-as-a-service. As shown in Fig. 1, the layers include infrastructure as a service, which provides the bottom layer of basic computing infrastructure; platform as a service; and software as a service.

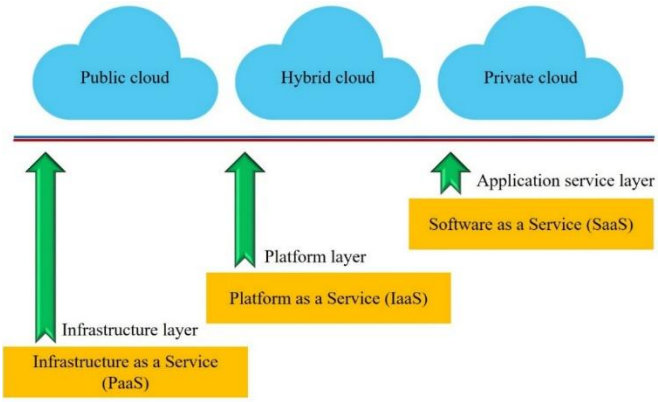


Fig. 1. Layered service architecture for cloud environments.

In practical cloud systems, task scheduling involves migrating computational workloads from user devices to large-scale cloud data centers. The objective is to ensure that tasks are processed efficiently while meeting user requirements for performance and service quality. Due to continuously changing workloads and resource conditions, scheduling mechanisms must be highly adaptable and capable of dynamically adjusting allocation strategies. Consequently, an effective scheduling system must consider factors such as system performance, resource constraints, task priorities, and user requirements to achieve efficient resource utilization. The architecture for service delivery in a cloud computing model is shown in Fig. 2 below. The architecture has several layers that process user requests and manage computational resources. At the user end, several clients simultaneously submit their work requests to the cloud computing system. These requests are handled by the task management layer.

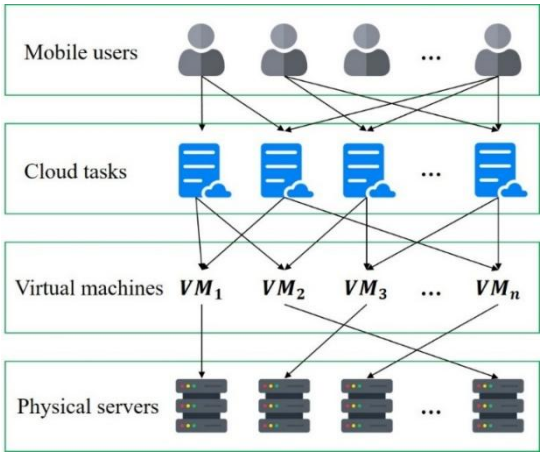


Fig. 2. Layered service delivery architecture in cloud computing.

To achieve this objective, the scheduling algorithm must consider these characteristics when allocating resources. The cloud computing architecture typically consists of multiple VMs running on physical computers. Computing resources can be allocated according to the system's needs. The resource manager keeps track of the current state of each computing resource, whether it is busy, idle, or faulty. Given the above facts and the chosen scheduling algorithm, the jobs are allocated to VMs. The performance analysis of the job, like its execution time, will be delivered to the user. Task scheduling is an important process in a cloud computing system. As shown in Fig. 3, the scheduling process typically comprises two main layers: the scheduling control layer and the task execution layer. First, users' submitted jobs are divided into individual tasks. Then, the scheduler evaluates all tasks to determine their relationships with computing resources. Resource allocation is then performed based on the task's needs.

Once the scheduling process is complete, the execution process is sent to the task-processing layer, where computing nodes perform the operations. While executing, the scheduler continuously monitors task status and adjusts allocations as needed. At the end of the execution process, the computation results are then returned to the user. Effective scheduling processes enable the cloud computing environment to work effectively.

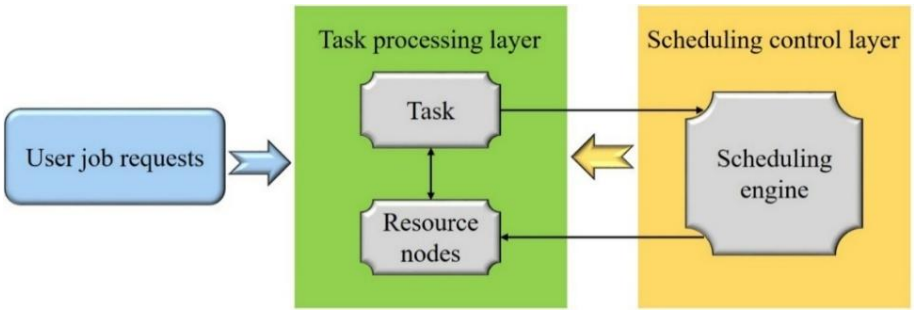


Fig. 3. Conceptual workflow of the cloud task scheduling mechanism.

B. Task and Resource Modeling

In cloud computing task scheduling optimization, the primary objective is to determine an optimal mapping between computational tasks and available computing resources. To formally describe the scheduling process, a mathematical model is constructed that captures the relationship between tasks and computing nodes. Let the cloud environment consist of a set of computing nodes $R = \{R_1, R_2, \dots, R_m\}$ and a set of tasks $T = \{T_1, T_2, \dots, T_n\}$, where m denotes the number of VMs, and n represents the number of submitted tasks, typically satisfying $n > m$. The scheduling decision can be represented using an assignment matrix as follows:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad (1)$$

where, a_{ij} is a binary decision variable indicating whether the task T_i is assigned to the resource R_j. If $a_{ij} = 1$, task T_i is executed on the j^{th} VM; otherwise, the value is zero. To accurately describe scheduling performance, three fundamental metrics are considered: makespan, resource utilization, and migration time. Makespan represents the completion time of the final task in the scheduling sequence, calculated using Eq. (2).

$$M = \max_{i \in \{1,\dots,m\}} C_i \quad (2)$$

where, C_i signifies the finishing time of the task T_i. Resource utilization quantifies the proportion of computational resources effectively used by scheduled tasks, defined as follows:

$$U = \frac{\sum_{j=1}^{n} \sum_{i=1}^{m} R_{ij} \cdot y_{ij}}{\sum_{j=1}^{n} R_j} \quad (3)$$

where, R_{ij} denotes the resource requirement of the task T_i on VM j, y_{ij} is a binary decision variable indicating whether the task T_i is assigned to VM j, and R_j represents the total capacity of VM j. Migration time captures the overhead incurred when tasks are transferred between VMs, expressed by Eq. (4).

$$T_{mig} = \sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij} \cdot o_{ij} \quad (4)$$

where, o_{ij} denotes the migration duration required for the task T_i to move to VM j. To evaluate scheduling quality, a composite fitness function is formulated by combining the three metrics. The overall optimization objective is defined by Eq. (5).

$$F = \omega_1 \left(\frac{1}{M}\right) + \omega_2 U - \omega_3 T_{mig} \quad (5)$$

where, ω_1, ω_2, and ω_3 represent the weighting coefficients that determine the importance of each objective.

IV. Proposed Algorithm

HBA operates on a population and mimics honey badger behavior when hunting in the wild. This algorithm consists of two main predation phases: the digging and honey phases. The latter allows the search technique to perform both exploration and exploitation to solve problems effectively. In nature, honey badgers collaborate with honey-guiding birds, as the latter are responsible for detecting bees' nests. Honey guides find the nest; however, honey badgers break open the hive with their claws. This cooperation forms the foundation for the algorithm. During the search process, potential solutions mimic honey badgers' prey-hunting behavior by scent-tracking and following clues.

A. Population Initialization

As with any other population-based algorithm, the search procedure begins with an initialization stage that generates candidate solutions. Let us assume that there are N_p honey badgers in the population. The location of the k^{th} badger is randomly generated from the search space using Eq. (6).

$$Y_k = L_b + (U_b - L_b) \cdot \rho, \quad k = 1,2, \dots, N_p \quad (6)$$

where, L_b and U_b represent the lower and upper bounds of the search space, respectively, while ρ is a uniformly distributed random number in the range [0,1]. The initialization process ensures that candidate solutions are uniformly distributed across the search space.

B. Digging Phase (Exploration)

During the digging phase, honey badgers detect prey using their powerful olfactory capabilities. This phase promotes global search, enabling the algorithm to find prey across extensive areas. The position updating equation used in this phase is given by Eq. (7).

$$y_i = y^* + G \cdot \kappa \cdot \Psi \cdot y^* + G \cdot r_2 \cdot \delta \cdot \Delta_i |\cos(2\pi r_3)[1 - \cos(2\pi r_4)]| \quad (7)$$

where, y^* represents the current optimal solution in the population, κ is a constant with a minimum value of 1, r_2, r_3,

and r_4 represent randomly generated numbers within [0,1], Δ_i denotes the distance between the current solution and the current optimal solution, and G serves as the directional parameter, regulating the search orientation. The digging phase includes an intensity operator that calculates the honey badger's attraction to the prey. This intensity is determined using Eq. (8).

$$\Psi = \rho_1 \cdot \frac{Q}{4\pi\Delta_i^2} \tag{8}$$

where, ρ_1 is a random number within [0,1] and Q represents the source strength. The source strength and distance term are computed as follows:

$$\mathrm{Q} = (y_i - y_{i+1})^2 \tag{9}$$

$$\Delta_i = y^* - y_i \tag{10}$$

To control the search direction, a directional control parameter is employed during the digging phase and is defined by Eq. (11).

$$G = \begin{cases} 1, & r_5 \le 0.5 \\ -1, & \text{otherwise} \end{cases} \tag{11}$$

where, r_5 represents a randomly generated number within the interval of [0,1]. If G is positive, the direction of movement will be towards the best solution; otherwise, it will be away from it. Moreover, a density parameter is used to control the gradual shift of the search mode from exploration to exploitation, decreasing nonlinearly with time and given by Eq. (12).

$$\delta = \theta \cdot \exp\left(-\frac{t}{T_{max}}\right) \tag{12}$$

where, θ is a constant usually set to 2, t is the current iteration number, and T_{max} denotes the maximum number of iterations. The adaptive parameter enables the algorithm to smoothly transition between global search and local refinement.

C. Honey Phase (Exploitation)

The next step of the algorithm models the interaction between the two species as cooperation. At this stage, the honey guides help locate the hive, and then the honey badger goes to the site to exploit the resources. The updating process at the honey stage can be mathematically defined by Eq. (13). It shows that the candidates approach the best-known location in a controlled random manner.

$$y_i = y^* + G \cdot r_6 \cdot \delta \cdot \Delta_i \tag{13}$$

D. Multi-Strategy Honey Badger Algorithm

While HBA shows excellent exploratory ability, it may be prone to an imbalance between exploration and exploitation. To address this drawback of the initial HBA, an improved approach, MHBA, will be introduced in this research. As the name suggests, this algorithm uses three improvement strategies: chaotic Bernoulli initialization, an optimal piecewise decrease strategy, and horizontal crossover. Chaotic maps produce deterministic sequences with high randomness, thereby enhancing population diversity during initialization. In the MHBA algorithm, instead of random initialization, the chaotic Bernoulli map is used. Eq. (14) defines the Bernoulli map.

$$z_{t+1} = \begin{cases} \frac{z_t}{1-\varepsilon}, & 0 \le z_t < 1-\varepsilon \\ \frac{z_t - (1-\varepsilon)}{\varepsilon}, & 1-\varepsilon \le z_t \le 1 \end{cases} \tag{14}$$

where, ε is a control parameter within the range [0,1]. A commonly used simplified form of the Bernoulli map is $z_{t+1} = 2z_t \bmod 1$ that generates a chaotic sequence used for initializing candidate solutions. To improve the adaptability of search behavior, MHBA employs the ODN technique. It controls the neighborhood search radius throughout the optimization process. The ODN parameter is expressed as:

$$ODN = \begin{cases} 1 + \tau_1 \cdot \eta_1 \cdot U(0,1), & 0 \le t \le \frac{T_{max}}{3} \\ 1 + \tau_2 \cdot \eta_2 \cdot U(0,1), & \frac{T_{max}}{3} < t \le \frac{2T_{max}}{3} \\ 1 + \tau_3 \cdot \eta_3 \cdot U(0,1), & \frac{2T_{max}}{3} < t \le T_{max} \end{cases} \tag{15}$$

Three nonlinear decline functions regulate the neighborhood size as follows:

$$w_1 = w_{max} - (w_{max} - w_{min})\left(\frac{t}{T_{max}}\right)^2 \tag{16}$$

$$w_2 = w_{max} - (w_{max} - w_{min})\frac{t}{T_{max}} \tag{17}$$

$$w_3 = w_{max} - (w_{max} - w_{min})\left[\frac{2t}{T_{max}} - \left(\frac{t}{T_{max}}\right)^2\right] \tag{18}$$

Early iterations emphasize exploration while later iterations focus on exploitation. By integrating ODN, the updated position in the digging phase becomes:

$$y_i = y^* \cdot ODN + G \cdot \kappa \cdot \Psi \cdot y^* + G \cdot r_2 \cdot \delta \cdot \Delta_i |\cos(2\pi r_3)(1 - \cos(2\pi r_4))| \tag{19}$$

Similarly, the honey stage is modified as follows:

$$y_i = y^* + ODN + G \cdot r_6 \cdot \delta \cdot \Delta_i \tag{20}$$

To enhance the exploration process, MHBA uses a crossover technique based on a swarm intelligence algorithm. This crossover creates new candidates by merging the two parents' information to create children's solutions using Eq.

(21) and (22).

$$y_{new}(i) = \begin{cases} c_1 y_i + (1-c_1)y_j + c_2(y_i - y_j), & \text{CSO/rand/1} \\ y_i + c_3(y_i - y_j), & \text{DE/rand/1} \\ y^* + c_4(y_i - y_j), & \text{DE/best/1} \\ y_i + c_5(y^* - y_i) + c_6(y_i - y_j), & DE/rand - to - best/1 \end{cases} \tag{21}$$

$$y_{new}(j) = \begin{cases} c_1 y_i + (1-c_1)y_j + c_2(y_j - y_i), & \text{CSO/rand/1} \\ y_j + c_3(y_j - y_i), & \text{DE/rand/1} \\ y^* + c_4(y_j - y_i), & \text{DE/best/1} \\ y_j + c_5(y^* - y_j) + c_6(y_j - y_i), & DE/rand - to - best/1 \end{cases} \quad (22)$$

The success ratio of each crossover technique is recorded, and the probability of selecting each technique is adaptively adjusted during the search process. The flowchart of the MHBA is depicted in Fig. 4. The MHBA algorithm begins by initializing a population using the Bernoulli chaotic mapping. During the search process, the algorithm alternates between the digging and honey stages using the PDN neighborhood approach. A horizontal crossover with adaptive selection of crossover techniques is employed to improve exploration. The search process stops when the stopping criterion is met; the optimal solution is yielded as the optimal schedule.

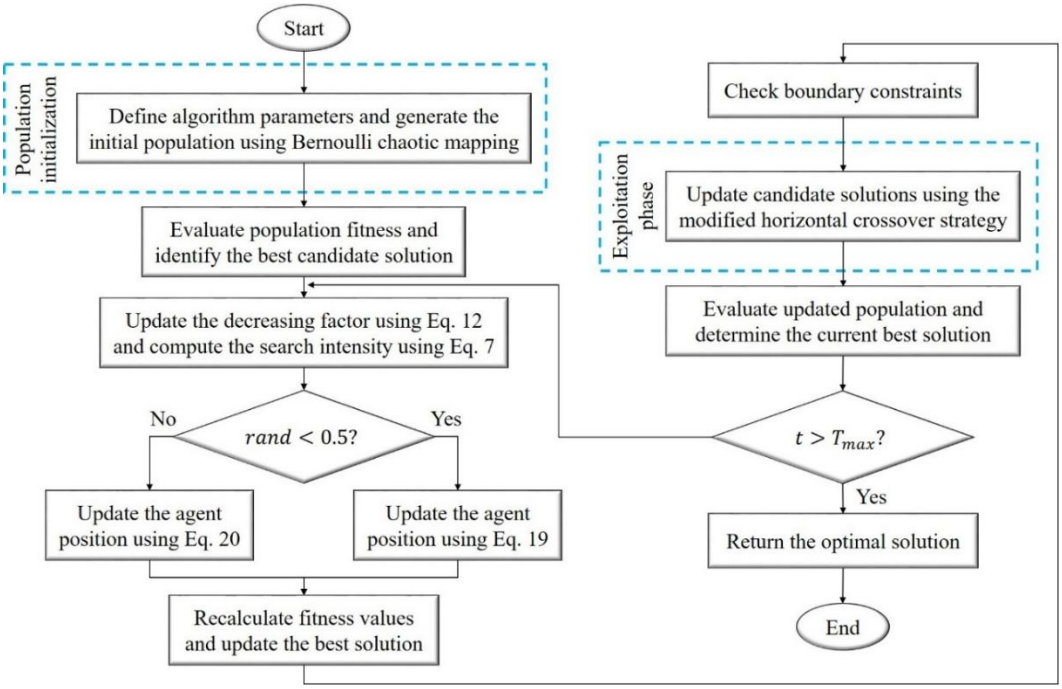


Fig. 4. Flowchart of the proposed algorithm.

V. Results and Discussion

To test the efficacy of the proposed MHBA for task scheduling in cloud computing, three critical parameters are used: makespan, resource utilization, and migration time. These parameters give a complete picture of how effective a task scheduling algorithm is in cloud computing. Makespan is defined as the total time required to execute all tasks scheduled in a system. A low makespan indicates that tasks are executed efficiently and that the task scheduling algorithm is effective. Resource utilization is the ratio of computational resources used to execute tasks to the total computational resources available in a system. If resource utilization is high, it indicates that a task scheduling algorithm is effective and that computational resources are being used efficiently in the system. Migration time is the total time required to migrate tasks from one VM to another within a system that executes tasks using a task scheduling algorithm. A low migration time implies that a task scheduling algorithm is effective.

The performance of the proposed MHBA-based scheduling framework is evaluated through simulation experiments using the CloudSim toolkit, which models and simulates the cloud computing environment. CloudSim provides a simulation framework for emulating complex cloud computing environments, including VMs, hosts, data centers, and task workloads, without requiring actual hardware. The simulation framework provides a wide range of cloud computing components, such as VM allocation, scheduling policies, and resource provisioning mechanisms. This provides a wide range of analysis capabilities for scheduling strategies in various cloud computing environments. To improve simulation reliability and realistic workload modeling, two workload traces, namely HPC2N and NASA, were used in simulation experiments. These workload traces contain detailed workload information regarding the execution behavior of jobs.

The HPC2N dataset is derived from a high-performance computing environment and comprises a heterogeneous workload analogous to large-scale scientific computing activities in a distributed computing environment. NASA is another dataset derived from scientific workflows executed in real computing environments. It comprises details on task execution time, task dependencies, and computational requirements. Using these datasets in the simulation enables the evaluation mechanism to capture realistic scheduling scenarios and workload diversity, which are common in cloud computing environments. By using the HPC2N and NASA datasets in the experiments, the MHBA algorithm is validated for its ability to handle diverse and dynamic task-scheduling scenarios. The configuration parameters utilized in the simulation environment for the cloud computing system are presented in Table II.

TABLE II. CONFIGURATION PARAMETERS USED IN THE CLOUD SIMULATION ENVIRONMENT

Simulation parameter	Assigned value
Total virtual machine instances	50
Processing capability of each VM	100 – 1000
Memory allocated per VM	1024 MB
Virtual machine memory capacity	9192 MB
Number of cloud data centers	2
Total number of submitted tasks	100 – 1000
Task workload size (instruction length)	1500 – 10,000
Virtualization platform	Xen

The makespan performance of MHBA has been tested against two popular swarm-based optimization techniques: PSO and Cat Swarm Optimization (CSO). Fig. 5 presents a visual representation of the results, and Table III summarizes the results for the HPC2N workload set. From 100 to 1000 cloudlets, all makespan values increase, indicating increased computational complexity. MHBA has been found to possess lower makespan values than the other techniques. When scheduling 100 cloudlets, MHBA achieves a makespan of 36.85, which is reduced by 25.5%, 34.57%, and 21.57% compared with PSO, CSO, and HBA, respectively. When the workload is increased to 500 cloudlets, MHBA still achieves the best performance, with a makespan of 135.18, which is reduced by 12.15%, 20.66%, and 8.9% compared with PSO, CSO, and HBA, respectively. When the workload is increased to 1000 cloudlets, MHBA still outperforms the other algorithms, achieving a makespan time of 247.25, which is reduced by 9.75%, 17.54%, and 7.34. Various parts of MHBA contribute to this improvement. For instance, the random initialization process enhances population diversity, ensuring the algorithm can explore all feasible scheduling options from the start.

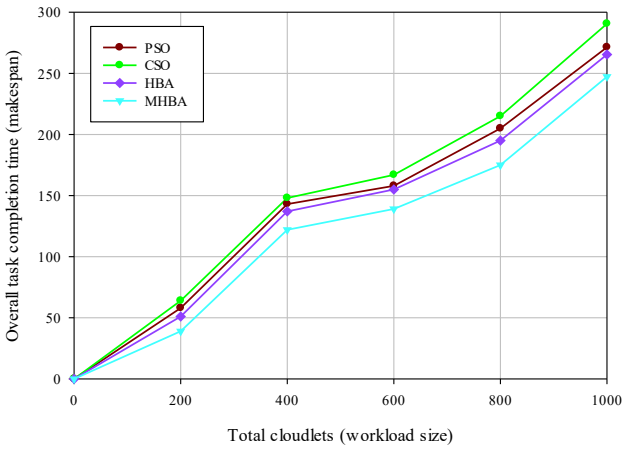


Fig. 5. Comparison of makespan for different scheduling algorithms under the HPC2N workload with varying numbers of cloud tasks.

The resource utilization comparison is shown in Fig. 6 and summarized in Table IV. As the number of tasks increases, the utilization rates of all algorithms decrease slightly. This is because the number of tasks is distributed among the available resources. Though MHBA has lower utilization for 100 cloudlets, the performance is still acceptable. This is because MHBA has strong exploration capabilities during the initial search phase. For smaller-scale tasks, the search space is relatively simple, and other algorithms, such as PSO and CSO, can achieve convergence.

TABLE III. COMPARISON OF MAKESPAN ACROSS SCHEDULING ALGORITHMS USING THE HPC2N DATASET

Workload size (Cloudlets)	PSO	CSO	HBA	MHBA
100	46.25	49.59	41.80	36.85
500	151.61	163.11	147.22	135.18
1000	271.38	290.63	265.42	247.25

MHBA has to spend some extra time exploring the possible scheduling configurations. MHBA outperforms other algorithms on complex tasks. For example, when the number of cloudlets is 500, MHBA has a better value, i.e., 2.42, for the resource utilization. This is because MHBA outperforms PSO, CSO, and HBA by 16.52%, 7.02%, and 8.67%, respectively. For 1000 cloudlets, MHBA outperforms other algorithms, achieving a resource utilization of 2.15. This is because MHBA outperforms PSO, CSO, and HBA by 16.74%, 13.95%, and 12.09%, respectively.

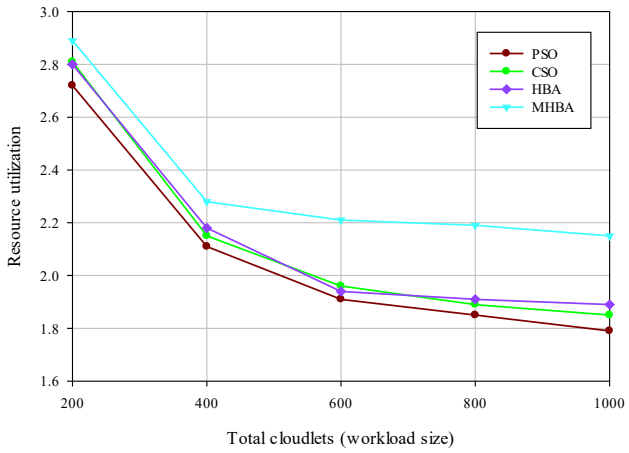


Fig. 6. Comparison of resource utilization for different scheduling algorithms under the HPC2N workload with varying numbers of cloud tasks.

TABLE IV. COMPARISON OF RESOURCE UTILIZATION ACROSS SCHEDULING ALGORITHMS USING THE HPC2N DATASET

Workload size (Cloudlets)	PSO	CSO	HBA	MHBA
100	2.94	3.03	2.97	3.15
500	2.02	2.25	2.21	2.42
1000	1.79	1.85	1.89	2.15

Fig. 7 and Table V show the migration time results. As the number of cloudlets increases (i.e., as system load increases), migration time increases for all algorithms. However, MHBA outperforms PSO, CSO, and HBA in terms of migration time. At 100 cloudlets, MHBA attains a migration time of 29.83, which is 28.76% lower than PSO, 38.08% lower than CSO, and 17.49% lower than HBA. At 500 cloudlets, MHBA attains 112.65, outperforming PSO and CSO by 11.2% and 20.6%, respectively. At 1000 cloudlets, MHBA again outperforms PSO, CSO, and HBA in terms of migration time, achieving 235.87, which is 15.16% lower than PSO, 27.89% lower than CSO, and 11.21% lower than HBA. This is because MHBA can identify

effective scheduling configurations early in the optimization process. Chaotic population initialization, dynamic search strategies, and strategy-based crossover in MHBA enable it to converge to effective task assignments early in the optimization process.

In addition to the HPC2N dataset, the proposed MHBA algorithm was further evaluated using the NASA workload traces. The performance results obtained using this dataset are shown in Fig. 8-10 and Table VI-VIII. Based on experimental analysis across various datasets, it is verified that the proposed MHBA algorithm achieves better scheduling efficiency across multiple performance metrics. In particular, the algorithm maintains a stable performance with heterogeneous tasks and varying resource requirements.

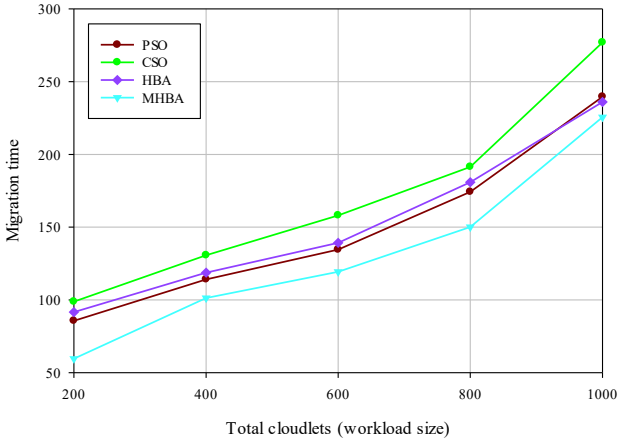


Fig. 7. Comparison of migration time for different scheduling algorithms under the HPC2N workload with varying numbers of cloud tasks.

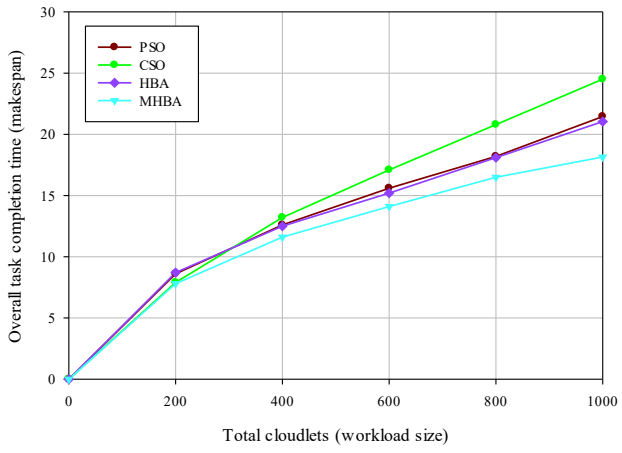


Fig. 8. Comparison of makespan for different scheduling algorithms under the NASA workload with varying numbers of cloud tasks.

TABLE V. COMPARISON OF MIGRATION TIME ACROSS SCHEDULING ALGORITHMS USING THE HPC2N DATASET

Workload size (Cloudlets)	PSO	CSO	HBA	MHBA
100	38.41	41.19	35.05	29.83
500	126.85	140.88	122.50	110.15
1000	239.69	277.01	236.12	225.66

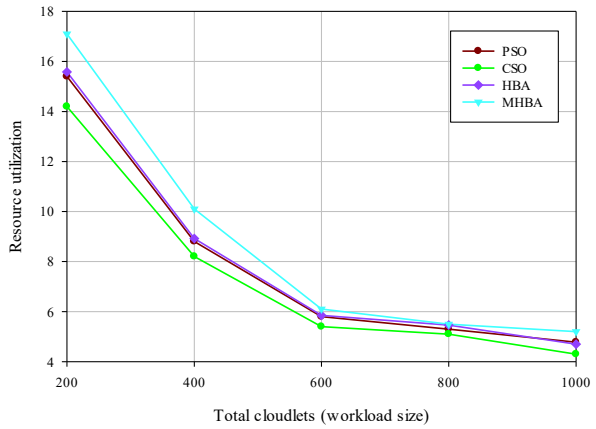


Fig. 9. Comparison of resource utilization for different scheduling algorithms under the NASA workload with varying numbers of cloud tasks.

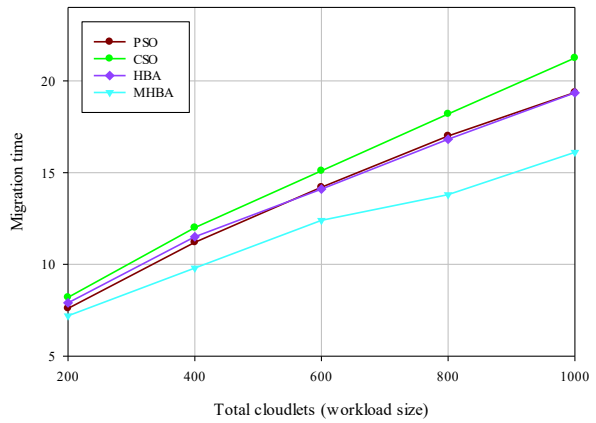


Fig. 10. Comparison of migration time for different scheduling algorithms under the NASA workload with varying numbers of cloud tasks.

TABLE VI. COMPARISON OF TASK MAKESPAN ACROSS SCHEDULING ALGORITHMS USING THE NASA DATASET

Workload size (Cloudlets)	PSO	CSO	HBA	MHBA
100	6.85	7.04	6.91	6.55
500	14.18	15.19	14.22	12.91
1000	21.45	24.51	21.05	18.15

TABLE VII. COMPARISON OF RESOURCE UTILIZATION ACROSS SCHEDULING ALGORITHMS USING THE NASA DATASET

Workload size (Cloudlets)	PSO	CSO	HBA	MHBA
100	18.42	16.80	18.85	21.45
500	6.18	6.25	6.31	6.61
1000	4.78	4.31	4.76	5.28

TABLE VIII. COMPARISON OF MIGRATION TIME ACROSS SCHEDULING ALGORITHMS USING THE NASA DATASET

Workload size (Cloudlets)	PSO	CSO	HBA	MHBA
100	5.92	6.78	5.98	5.71
500	13.06	13.67	13.02	11.20
1000	19.37	21.25	19.35	16.11

VI. CONCLUSION

In this study, a novel algorithm, MHBA, was proposed for the multi-target task scheduling issue in a cloud computing context. Firstly, a chaotic Bernoulli initialization strategy was adopted to improve the diversity of the initial population and expand the search area. In this way, the algorithm's randomness is enhanced, and its convergence is improved. The second contribution of this study is the development of a piecewise optimum decreasing neighborhood model. In this way, the algorithm's search space is optimized, thereby improving its convergence to the solution. The efficiency of the MHBA algorithm has been validated using experimental results. It can be observed that, compared with the other baseline algorithms, MHBA outperforms them across various metrics at different workloads. The advantages of MHBA become more pronounced as the workload increases. MHBA is better suited to large-scale scheduling problems. In addition, the significance test results indicate that MHBA's advantages are statistically significant.

MHBA's better performance can be attributed to the synergy among its three enhancement strategies. The chaotic initialization mechanism enhances the algorithm's global search, the adaptive neighborhood strategy balances exploration and exploitation, and the horizontal crossover mechanism improves population diversity and information sharing. These mechanisms enable MHBA to have outstanding optimization ability. As for future research, several avenues are outlined. Firstly, it is possible to propose adaptive parameter control mechanisms that can improve the algorithm's robustness in dynamic environments. Secondly, it is possible to incorporate energy/cost constraints into the scheduling process, thereby improving the algorithm's applicability to green computing. Thirdly, it is possible to extend the MHBA framework to collaborative environments and real-time task scheduling, and to improve the algorithm's applicability to emerging/developing distributed computing environments, e.g., the Internet of Things and edge computing.

REFERENCES

[1] Q. Li and S. Lang, "Dynamic resource mapping in cloud computing workflows using the modified artificial fish swarm algorithm with dynamic vision and step-size adaptation," *Computing,* vol. 108, no. 4, p. 59, 2026.

[2] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single‐objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," *Concurrency and Computation: Practice and Experience,* vol. 34, no. 5, p. e6698, 2022.

[3] T. Hagras and G. A. El-Sayed, "A fault-tolerant and load-balancing scheduler for independent tasks on cloud-based virtual machines," *Cluster Computing,* vol. 29, no. 1, p. 61, 2026.

[4] A. Ayyadi and A. Jahani, "Addressing cost and resource variability for big data task scheduling in heterogeneous cloud environments: A. Ayyadi, A. Jahani," *The Journal of Supercomputing,* vol. 81, no. 6, p. 762, 2025.

[5] Q. Jiang, X. Xin, T. Zhang, and K. Chen, "Energy-Efficient Task Scheduling and Resource Allocation in Edge Heterogeneous Computing Systems Using Multi-Objective Optimization," *IEEE Internet of Things Journal,* 2025.

[6] K. Danach, K. Hamze, H. Harb, and H. Kanj, "Carbon‐Aware Scheduling in Cloud Computing Operations: A Multi‐Objective Optimisation Approach," *IET Smart Grid,* vol. 9, no. 1, p. e70056, 2026.

[7] F. Safi-Esfahani, H. Larian, S. Saeedi Mobarakeh, and S. Mirjalili, "Dynamic scheduling of independent tasks in cloud computing environment applying improved chicken swarm optimization and differential evolution," *The Journal of Supercomputing,* vol. 81, no. 8, p. 867, 2025.

[8] B. Hariharan, C. Anupama, R. K. Neerukonda, and R. K. Dhanaraj, "Energy‐Aware System for Dynamic Workflow Scheduling in Cloud Data Centers: A Genetic Algorithm with DQN Approach," *Networked Sensing Systems,* pp. 121–143, 2025.

[9] B. Mondal, "Optimization techniques for big data: A VM load balancing in cloud computing using particle swarm optimization," in *Mathematical Modeling for Big Data Analytics*: Elsevier, 2026, pp. 143–156.

[10] S. Roy, R. Acharya, C. Chowdhury, S. Chattopadhyay, and P. Bhaumik, "An ant colony optimization based job allocation strategy for IoT-edge-cloud computing environment," *Journal of Ambient Intelligence and Humanized Computing,* vol. 17, no. 1, pp. 125–139, 2026.

[11] F. A. Hashim, E. H. Houssein, K. Hussain, M. S. Mabrouk, and W. Al-Atabany, "Honey Badger Algorithm: New metaheuristic algorithm for solving optimization problems," *Mathematics and Computers in Simulation,* vol. 192, pp. 84–110, 2022.

[12] P. Pirozmand, H. Jalalinejad, A. A. R. Hosseinabadi, S. Mirkamali, and Y. Li, "An improved particle swarm optimization algorithm for task scheduling in cloud computing," *Journal of Ambient Intelligence and Humanized Computing,* vol. 14, no. 4, pp. 4313–4327, 2023.

[13] R. Gong, D. Li, L. Hong, and N. Xie, "Task scheduling in cloud computing environment based on enhanced marine predator algorithm," *Cluster Computing,* vol. 27, no. 1, pp. 1109–1123, 2024.

[14] P. Pabitha, K. Nivitha, C. Gunavathi, and B. Panjavarnam, "A chameleon and remora search optimization algorithm for handling task scheduling uncertainty problem in cloud computing," *Sustainable Computing: Informatics and Systems,* vol. 41, p. 100944, 2024.

[15] M. Selselejoo and H. Ahmadifar, "DT-GWO: A Hybrid Decision Tree and GWO-Based Algorithm for Multi-Objective Task Scheduling Optimization in Cloud Computing," *Sustainable Computing: Informatics and Systems,* p. 101138, 2025.

[16] F. A. Emara, A. A. Gad-Elrab, A. Sobhi, A. S. Alsharkawy, M. E. Embabi, and M. Abd El-Baky, "Multi-objective task scheduling algorithm for load balancing in cloud computing based on improved Harris hawks optimization," *The Journal of Supercomputing,* vol. 81, no. 6, pp. 1–38, 2025.

[17] L. Du, T. Xie, and B. Chen, "Adaptive chaotic reverse learning-enhanced reptile search algorithm for efficient task scheduling in cloud computing," *The Journal of Supercomputing,* vol. 82, no. 2, p. 105, 2026.

[18] A. B. Kathole, A. S. Mirge, and S. G. Patil, "Optimizing cloud task scheduling and VM placement with multi-objective constraints using hybrid lemurs and gannet optimization algorithm," *Multimedia Tools and Applications,* vol. 85, no. 2, p. 107, 2026.

[19] S. Awad, M. Gamal, K. A. El Salam, and R. F. Abdel-Kader, "Efficient workflow scheduling in fog-cloud collaboration using a hybrid IPSO-GWO algorithm," *Scientific Reports,* vol. 16, no. 1, p. 4359, 2026.

[20] M. A. Tawfeek, I. Alrashdi, M. Alruwaili, B. Aldughayfiq, and H. Allahem, "An enhanced hybrid artificial bee colony and genetic algorithm for multi-objective workflow scheduling in the cloud," *Computing,* vol. 108, no. 2, p. 28, 2026.

[21] B. Rajeshwari and M. Namratha, "A Novel Push Pull Optimization Algorithm Fostered Deadline Aware Task Scheduling in Virtualized Cloud Computing," *Franklin Open,* p. 100499, 2026.

Structure-Aware Latent Diffusion for High-Quality Line Art Colorization

Shuhua Xu[1], Qiang Ai[2], An Zhao[3], Guan Yang[4]*, Bo Chen[5]
School of Intelligent Clothing and Apparel, Zhongyuan University of Technology, Zhengzhou, China[1]
School of Information Science and Engineering, Lanzhou University, Lanzhou, China[2]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, China[3]
School of Intelligent Perception and Instrumentation, Zhongyuan University of Technology, Zhengzhou, China[4]
School of Mathematical Sciences, Shenzhen University, Shenzhen, China[5]
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen, China[5]

***Abstract*—To address the limitations of existing line art colorization methods in structural preservation, color mapping accuracy, and semantic consistency, this study proposes a structure-aware multi-instance constrained line art colorization method based on latent diffusion. Built upon the latent diffusion framework, the proposed method introduces a structure-aware constraint mechanism to enhance the preservation of line contours and edge details during generation. Meanwhile, instance-level semantic modeling and feature fusion strategies are incorporated to achieve coherent local color representation and optimize global semantic consistency. In addition, a unified optimization objective is constructed by jointly integrating structural constraints, color consistency constraints, and regularization terms, thereby improving the visual quality and naturalness of the generated results through collaborative multi-constraint learning. Experimental results on public datasets demonstrate that the proposed method outperforms comparative approaches in terms of FID, PSNR, SSIM, and LPIPS, producing high-quality colorization results with clear structures, natural colors, and strong semantic consistency, which verifies its effectiveness and superiority.**

***Keywords*—Line art colorization; latent diffusion; structure-aware modeling; instance-level semantic modeling; feature fusion; image generation**

I. Introduction

In modern electronic measurement and intelligent vision systems, high-precision image acquisition and processing play a crucial role in automated inspection, analysis, and visualization tasks [1], [2]. In applications such as electronic component detection, industrial visual monitoring, and intelligent measurement instruments, the accurate interpretation of image information directly affects the reliability and precision of measurement results [3], [4]. In recent years, with the widespread adoption of animation-style visual content and line art representations in education, electronic displays, digital media, and intelligent human–computer interaction systems, efficient automatic line art colorization has gradually emerged as an important research topic in image processing and electronic information systems [5], [6].

Traditional manual colorization methods are labor-intensive and time-consuming, and they often suffer from inconsistent color representation and structural ambiguity when handling multiple object instances [7], [8], [9]. These limitations can significantly degrade visualization quality and affect observation accuracy in practical applications [10]. With the rapid development of deep generative models, diffusion models have demonstrated remarkable performance in image synthesis and restoration tasks owing to their strong generative capability and latent-space representation learning ability [5], [11]. Nevertheless, existing diffusion-based line art colorization methods still face several challenges in multi-instance scenarios, including insufficient instance-level semantic understanding, inadequate exploitation of structural contour information, and excessive reliance on complex conditional guidance strategies in latent space [12], [13].

To address the above issues, this study proposes a structure-constrained multi-instance line art colorization framework based on latent diffusion, termed SC-MILC (*Structure-Constrained Multi-Instance Line Art Colorization*). Specifically, the proposed framework consists of three key components: a latent generative backbone (LGB), a structure-aware guidance module (SGM), and an instance semantic modeling and feature fusion module (ISMFF). The latent generative backbone performs efficient color generation in latent space, while the structure-aware guidance module enhances contour preservation and structural consistency by exploiting line art edge information. Furthermore, the instance semantic modeling and feature fusion module enables accurate semantic representation and adaptive feature interaction among multiple instances. On this basis, a unified optimization objective is constructed to jointly model diffusion reconstruction loss, structural consistency constraints, color consistency supervision, and instance fusion regularization, thereby achieving high-quality and controllable multi-instance line art colorization.

The main contributions of this work are summarized as follows:

- A latent diffusion-based generation framework for multi-instance line art colorization is proposed, which improves generation stability and computational efficiency.
- A structure-aware guidance module is introduced to enhance structural consistency and edge fidelity by effectively utilizing line art contour information.
- An instance semantic modeling and gated feature fu-

*Corresponding author

sion mechanism is designed to achieve accurate multi-instance color control while balancing local detail representation and global semantic consistency.

- A unified optimization objective function is developed to provide an effective loss evaluation strategy for latent-space multi-instance line art colorization.

II. Problem Definition

The objective of multi-instance line art colorization is to transform an input line art image into a high-quality color image while achieving accurate color control under multiple object instances. Let the input line art image be defined as:

$$S \in \mathbb{R}^{H \times W}, \tag{1}$$

where, H and W denote the height and width of the image, respectively. The reference instance set is represented as:

$$\mathcal{R} = \{I_i\}_{i=1}^{N}, \quad I_i \in \mathbb{R}^{H \times W \times 3}, \tag{2}$$

where, N denotes the number of object instances and each reference image I_i provides the target color information for the corresponding instance. The spatial mask set associated with the reference instances is defined as:

$$\mathcal{M} = \{M_i\}_{i=1}^{N}, \quad M_i \in \{0,1\}^{H \times W}, \tag{3}$$

where, M_i is a binary mask indicating the spatial region occupied by the i-th instance. The goal of the proposed framework is to learn a conditional mapping function:

$$\mathcal{F} : (S, \mathcal{R}, \mathcal{M}) \mapsto \hat{X}, \quad \hat{X} \in \mathbb{R}^{H \times W \times 3}, \tag{4}$$

such that the generated image $\hat{X}$ satisfies the following constraints.

A. Structural Consistency

The generated result should preserve the contour structures and geometric layout of the input line art image, which can be formulated as:

$$\mathcal{F}_{edge}(\hat{X}) \approx \mathcal{F}_{edge}(S), \tag{5}$$

where, $\mathcal{F}_{edge}(\cdot)$ denotes the edge extraction operator.

B. Instance-Level Color Consistency

For each instance region M_i, the generated colors should remain consistent with those of the corresponding reference instance I_i, which is expressed as:

$$\mathrm{Pool}_{M_i}(\hat{X}) \approx \mathrm{Pool}_{M_i}(I_i), \tag{6}$$

where, $\mathrm{Pool}_{M_i}(\cdot)$ represents the average pooling or feature extraction operation within the masked region M_i.

C. Multi-Instance Semantic Consistency

Under multi-instance conditions, the framework should maintain a balance between local instance-level representations and global semantic information, thereby ensuring natural color distributions and avoiding color conflicts among different instances.

Based on the above constraints, the entire task can be formulated as a conditional generation problem in latent space. Let the latent representation of the input line art image be:

$$z_0 = \mathcal{E}(S), \tag{7}$$

where, $\mathcal{E}(\cdot)$ denotes the encoder. The final generated image is obtained by:

$$\hat{X} = \mathcal{D}\left(\mathcal{F}_{diff}\left(z_0, \{f_i, M_i\}_{i=1}^{N}, F_{struct}\right)\right), \tag{8}$$

where, $\mathcal{D}(\cdot)$ denotes the decoder, $\mathcal{F}_{diff}$ represents the latent diffusion generation backbone, f_i denotes the semantic feature of the i-th instance, and F_{struct} represents the structure-aware feature representation.

Through the above formulation, the multi-instance line art colorization task can be uniformly modeled as an optimization problem under a latent conditional generation framework, jointly constrained by structural consistency, instance-level color supervision, and semantic coherence. This formulation provides a unified theoretical foundation for the subsequent network architecture and optimization strategy design.

III. Methodology

The overall architecture of the proposed SC-MILC framework is illustrated in Fig. 1. The framework consists of four major components: a latent-space generation backbone, a structure-aware guidance branch, an instance semantic modeling module, and a reference feature fusion module. The framework takes a line art image together with multiple reference instance images as conditional inputs and performs progressive denoising generation in latent space, thereby achieving automatic multi-instance colorization and consistent color control.

Specifically, a pre-trained variational autoencoder (VAE) is first employed to project the target image into a low-dimensional latent space, where the diffusion process is conducted to reduce computational complexity and memory consumption. Subsequently, the input line art image is fed into the structure-aware guidance branch to extract contour and edge representations. These structural features are further utilized to construct structure-aware constraints, which enhance spatial consistency during the generation process.

Meanwhile, multiple reference instance images are processed by the semantic encoding module to extract high-level semantic representations. Combined with the corresponding instance region masks, the extracted features form region-aware reference representations with instance-level perception capability. Based on these representations, the reference feature fusion module adaptively integrates global semantic information and local instance features to generate unified conditional control signals.

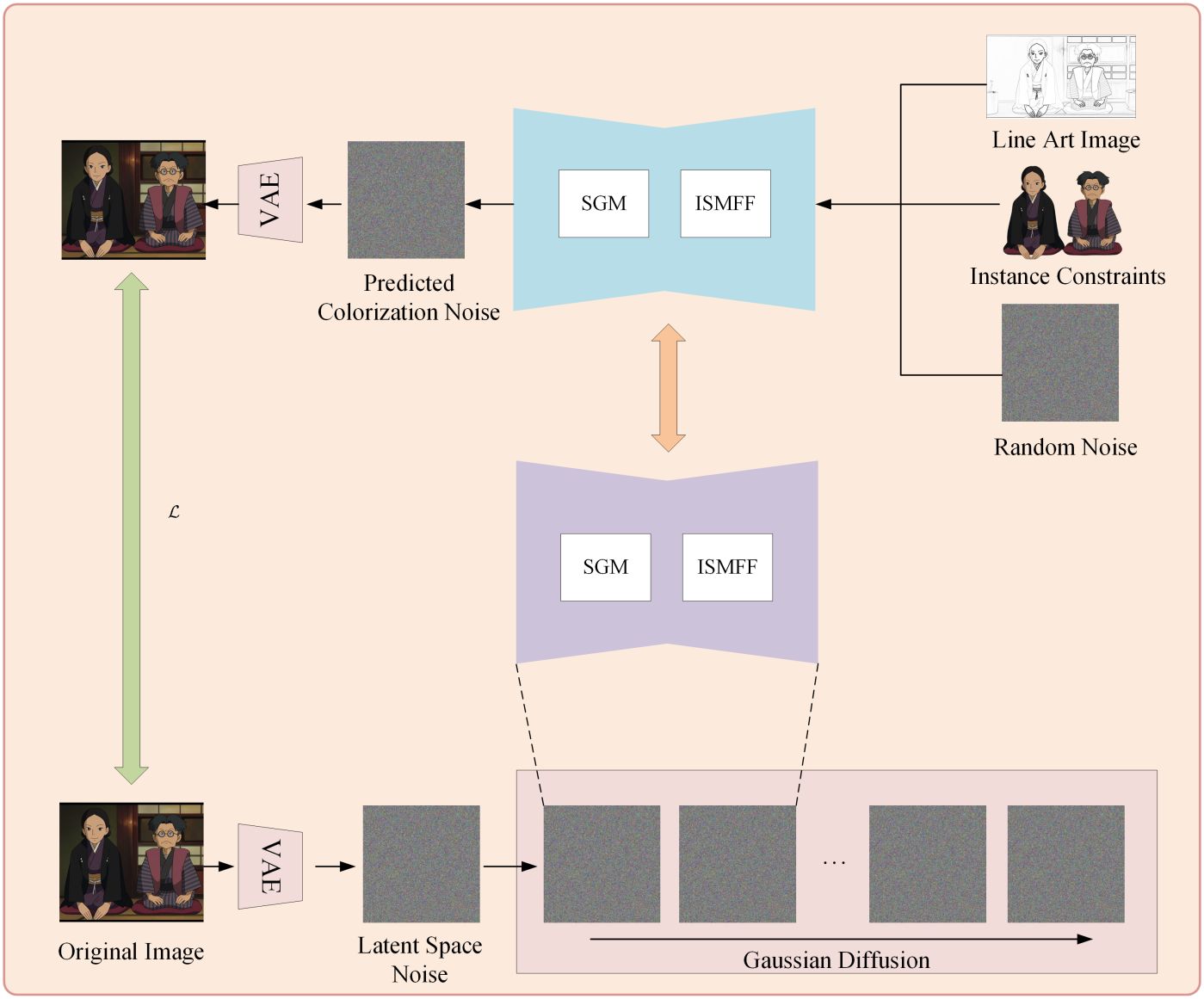


Fig. 1. Overall architecture of the proposed SC-MILC framework. The framework consists of a latent-space diffusion generation backbone, a structure-aware guidance branch, an instance semantic modeling module, and a reference feature fusion module for multi-instance line art colorization.

During the diffusion denoising process, latent features are progressively restored into clear representations under the joint constraints of multiple conditional signals. Among them, structural information constrains the spatial layout and contour consistency of the generated results, while reference semantic information guides color distributions and texture details. Finally, the generated latent representations are decoded into color images through the decoder, producing high-quality multi-instance line art colorization results.

A. Latent Generative Backbone (LGB)

To achieve efficient image generation modeling, a latent-space diffusion generation backbone is constructed in the proposed framework. Let the input image be denoted as $x \in \mathbb{R}^{H\times W\times 3}$. A pre-trained variational autoencoder (VAE) encoder $\mathcal{E}(\cdot)$ is first employed to project the image into a compact latent representation:

$$z_0 = \mathcal{E}(x), \quad z_0 \in \mathbb{R}^{C\times H'\times W'}, \tag{9}$$

where, $H' = \frac{H}{s}$ and $W' = \frac{W}{s}$ denote the latent spatial resolution, and s represents the spatial downsampling factor.

In the latent space, a forward diffusion process is defined by progressively injecting Gaussian noise into z_0 to construct a sequence of latent variables $\{z_t\}_{t=1}^{T}$,

$$q(z_t|z_0) = \mathcal{N}\left(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{10}$$

where, $\alpha_t = 1-\beta_t$, $\beta_t \in (0,1)$ denotes the predefined noise scheduling parameter, and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. The above process can be equivalently reformulated as:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}), \tag{11}$$

which enables direct sampling of noisy latent representations at arbitrary diffusion steps.

Based on the above formulation, a parameterized noise prediction model ϵ_θ is introduced to estimate the injected noise and progressively recover the latent distribution through the reverse diffusion process. The conditional reverse transition probability is defined as:

$$p_\theta(z_{t-1}|z_t, c) = \mathcal{N}\left(z_{t-1}; \mu_\theta(z_t, t, c), \sigma_t^2 \mathbf{I}\right), \tag{12}$$

where, c denotes external conditional information, and the mean parameter μ_θ is estimated by the noise prediction network:

$$\mu_\theta(z_t, t, c) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, c)\right). \tag{13}$$

Consequently, the latent-space generation process can be uniformly formulated as a learning problem for the noise prediction function $\epsilon_\theta(z_t, t, c)$. The optimization objective is constructed by minimizing the mean squared error between the predicted noise and the ground-truth Gaussian noise:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]. \tag{14}$$

Under the above modeling framework, the latent variables progressively converge from the noise distribution toward the latent representation corresponding to the real data distribution under the guidance of conditional information c. Finally, the generated latent representation is mapped back to the image space through the decoder $\mathcal{D}(\cdot)$:

$$\hat{x} = \mathcal{D}(z_0). \tag{15}$$

The proposed latent generative backbone provides a compact and unified formulation for modeling the transformation from the data distribution to the noise distribution and subsequently to the generated distribution. This formulation further establishes a unified conditional generation foundation for integrating structural guidance and instance-level semantic information in the subsequent modules.

B. Structure-Aware Guidance Module (SGM)

To enhance the structural consistency of the generated results, a structure-aware guidance mechanism is introduced to explicitly incorporate geometric contour information from line art images into the diffusion modeling process. Let the input line art image be denoted as $S \in \mathbb{R}^{H\times W}$. Its structural response is first extracted using an edge operator:

$$E = \mathcal{F}_{edge}(S), \tag{16}$$

where, $\mathcal{F}_{edge}(\cdot)$ denotes the structure extraction operator, and $E \in \mathbb{R}^{H\times W}$ represents the normalized edge intensity map.

To align the structural representation with latent-space features, the extracted edge map is further projected to the latent spatial resolution:

$$E' = \mathcal{R}(E), \quad E' \in \mathbb{R}^{H'\times W'}, \tag{17}$$

where, $\mathcal{R}(\cdot)$ denotes the spatial rescaling operation.

During the diffusion denoising process, let the intermediate latent feature representation be denoted as $F_t \in \mathbb{R}^{C\times H'\times W'}$. The conventional attention operation is formulated as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{18}$$

where, Q, K, and V denote the query, key, and value feature mappings, respectively, and d represents the feature dimension.

To introduce structural constraints into the attention mechanism, a structure-aware modulation term A_s is constructed as:

$$A_s = \gamma \cdot \mathrm{Norm}(E'), \tag{19}$$

where, $\mathrm{Norm}(\cdot)$ denotes the normalization operation and γ is a learnable modulation parameter.

The structure-aware attention operation is then formulated as:

$$\mathrm{Attn}_s(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + A_s\right)V, \tag{20}$$

which enables the attention distribution to adaptively focus on structurally important regions.

Furthermore, to strengthen feature representation in edge regions, an element-wise modulation strategy is introduced for intermediate latent features:

$$\tilde{F}_t = F_t \odot (1 + \alpha E'), \tag{21}$$

where, $\odot$ denotes element-wise multiplication and α represents the modulation coefficient. This operation assigns higher response weights to structural edge regions, thereby suppressing color overflow and structural ambiguity during generation.

Through the above mechanisms, structural information is jointly incorporated into the diffusion process via attention bias modulation and feature-level enhancement. Consequently, the proposed structure-aware guidance module enables the diffusion model to preserve clear structural boundaries and stable spatial layouts during latent-space denoising, thereby improving the structural consistency between the input line art image and the generated colorized result.

C. Instance Semantic Modeling and Feature Fusion Module (ISMFF)

To achieve accurate color control under multi-instance conditions, a unified instance semantic modeling and feature fusion mechanism is constructed in the proposed framework. Let the reference image set be denoted as $\mathcal{R} = \{I_i\}_{i=1}^{N}$, and the corresponding instance mask set be represented as $\mathcal{M} = \{M_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$ and $M_i \in \{0,1\}^{H\times W}$.

First, each reference instance image is processed by a semantic encoder $\mathcal{F}_{enc}(\cdot)$ to extract high-level semantic representations:

$$f_i = \mathcal{F}_{enc}(I_i), \quad f_i \in \mathbb{R}^{C\times H'\times W'}, \tag{22}$$

where, f_i denotes the semantic feature representation of the i-th reference instance.

To establish spatial correspondence between semantic features and instance regions, the instance masks are projected to the latent-space resolution:

$$M_i' = \mathcal{R}(M_i), \quad M_i' \in \mathbb{R}^{H'\times W'}, \tag{23}$$

where, $\mathcal{R}(\cdot)$ denotes the spatial rescaling operation.

Based on the projected masks, instance-aware semantic features are defined as:

$$\hat{f}_i = f_i \odot M_i', \tag{24}$$

where, $\odot$ denotes element-wise multiplication. This operation explicitly constrains semantic features within the corresponding instance regions.

To further enhance global semantic perception, a global semantic representation is introduced as:

$$f_g = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Pool}(f_i), \tag{25}$$

where, $\mathrm{Pool}(\cdot)$ denotes the global pooling operation.

Furthermore, an adaptive feature fusion mechanism is designed to integrate local instance features with global semantic information. The gating function is formulated as:

$$g_i = \sigma\left(W_1 f_g + W_2 \mathrm{Pool}(\hat{f}_i)\right), \tag{26}$$

where, $\sigma(\cdot)$ denotes the Sigmoid activation function, and W_1 and W_2 represent learnable parameters.

Based on the obtained gating weights, the fused instance feature representation is defined as:

$$\tilde{f}_i = g_i \cdot \hat{f}_i + (1 - g_i)\cdot f_g, \tag{27}$$

which enables adaptive balancing between local instance-specific information and global semantic consistency.

Finally, all fused instance features are aggregated to construct a unified instance-aware conditional representation:

$$F_{inst} = \sum_{i=1}^{N} \tilde{f}_i, \tag{28}$$

which is incorporated into the diffusion generation framework as part of the conditional guidance signal:

$$c = \{F_{inst}, F_{struct}\}, \tag{29}$$

where, F_{struct} denotes the structure-aware feature representation generated by the structure-aware guidance module.

Through the above formulation, instance-level semantic alignment is explicitly achieved by introducing mask-based spatial constraints, while the proposed gating mechanism adaptively balances local instance features and global semantic information. Consequently, the proposed ISMFF module enables stable and semantically consistent color mapping in multi-instance scenarios. Compared with conventional methods relying on strict spatial feature alignment, the proposed strategy effectively reduces feature alignment complexity while improving semantic representation capability and model generalization performance.

D. Optimization Objective and Constraint Functions

Under the unified framework of the latent generative backbone, the structure-aware guidance module, and the instance semantic fusion module, an overall optimization objective is constructed to ensure structural consistency and color fidelity in the generated images.

Let $\epsilon_\theta(z_t, t, c)$ denote the predicted noise in latent space and ϵ represent the ground-truth Gaussian noise. The basic diffusion optimization objective is defined as:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]. \tag{30}$$

To enhance structural fidelity, a structure-aware loss based on edge consistency is introduced:

$$\mathcal{L}_{struct} = \mathbb{E}_{z_0}\left[\|\mathcal{F}_{edge}(\mathcal{D}(z_0)) - E\|_2^2\right], \tag{31}$$

where, $\mathcal{D}(\cdot)$ denotes the latent-space decoder, E represents the edge map extracted from the input line art image, and $\mathcal{F}_{edge}(\cdot)$ denotes the edge extraction operator.

To achieve instance-level color consistency, an instance-aware color matching loss is further defined as:

$$\mathcal{L}_{color} = \frac{1}{N}\sum_{i=1}^{N} \|\mathrm{Pool}_{M_i}(\mathcal{D}(z_0)) - \mathrm{Pool}_{M_i}(I_i)\|_2^2, \tag{32}$$

where, $\mathrm{Pool}_{M_i}(\cdot)$ denotes masked average feature extraction within the instance region specified by M_i, and I_i denotes the corresponding reference instance image.

To prevent the gating mechanism from collapsing toward either the local instance branch or the global semantic branch, a gated regularization term is introduced:

$$\mathcal{L}_{gate} = \frac{1}{N}\sum_{i=1}^{N} \|g_i - 0.5\|_2^2, \tag{33}$$

which encourages the gating coefficients to maintain a balanced contribution from both local instance features and global semantic representations during feature fusion. By avoiding extreme gate values, the proposed regularization improves the stability of feature integration and facilitates more effective exploitation of complementary information across different semantic levels.

By combining the above objectives, the final overall optimization function is formulated as:

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_s \mathcal{L}_{struct} + \lambda_c \mathcal{L}_{color} + \lambda_g \mathcal{L}_{gate}, \tag{34}$$

where, λ_s, λ_c, and λ_g denote balancing hyperparameters used to control the contributions of structural consistency, color supervision, and feature fusion regularization, respectively.

Through the above optimization strategy, the proposed framework enables the diffusion model to simultaneously satisfy the following objectives during latent-space denoising: preserving clear structural contours and spatial layouts from the input line art image, maintaining color consistency between generated instances and reference images, and balancing local instance semantics with global contextual representations to produce visually natural and semantically coherent colorization results.

IV. Dataset and Preprocessing

In this study, two publicly available anime line art datasets, *Sakuga* [14] and *ATD-12K* [15], are adopted to evaluate the performance of the proposed framework on multi-instance line art colorization tasks.

The Sakuga dataset contains high-quality animation frame sequences with rich motion variations and diverse character poses, making it suitable for training and evaluating multi-instance colorization models. In contrast, the ATD-12K dataset provides well-annotated line art and color image pairs covering various character categories and artistic styles, thereby offering reliable references for instance-level color learning.

As illustrated in Fig. 2, both datasets are uniformly processed to construct the final training and testing sets. During preprocessing, samples with excessively low resolution or incomplete line art structures are removed to ensure data quality. For each image pair, the line art image S and the corresponding color reference image I are jointly utilized to establish the conditional colorization relationship. Furthermore, instance masks are extracted as auxiliary conditional inputs for multi-instance semantic modeling.

Specifically, an open-source segmentation tool is first employed to segment characters and key objects in the reference images, thereby generating binary instance masks to support instance-level semantic modeling. Subsequently, both the line art images and the color reference images are uniformly resized to a fixed spatial resolution while preserving the original aspect ratio to maintain latent-space feature alignment.

Fig. 2. Dataset preprocessing and instance mask construction pipeline. The line art images, reference color images, and corresponding instance masks are jointly utilized for multi-instance conditional colorization training.

To improve the generalization capability of the model under varying viewpoints and illumination conditions, data augmentation strategies including random horizontal flipping and brightness perturbation are further applied to the reference images. These augmentation operations effectively enhance the robustness of the proposed framework in modeling instance-level color distributions and structural representations.

The above dataset partitioning and preprocessing strategies ensure the reproducibility of the experiments and provide sufficient multi-instance conditional information for validating the effectiveness of the latent generative backbone, the structure-aware guidance module, and the instance semantic fusion mechanism.

A. Experimental Environment and Parameter Settings

All experiments were implemented using the PyTorch deep learning framework. The detailed experimental environment configuration is summarized in Table I.

TABLE I. Experimental Environment Settings

Environment	Configuration
CPU	Xeon(R) Gold 6348
GPU	NVIDIA A800 80GB
Memory	120 GB
Programming Language	Python 3.9
Deep Learning Framework	PyTorch 2.6.0 + cu118

During the training stage, in order to ensure the stability and convergence of the latent generative backbone and the instance semantic fusion module, the proposed framework was trained for a total of 100,000 optimization iterations. A batch size of 1 was adopted for single-sample training to fully preserve instance-level conditional information during the diffusion learning process.

The initial learning rate was set to 1×10^{-4} and remained fixed throughout the training process to maintain stable latent-space noise prediction and conditional feature optimization. The AdamW optimizer was employed for parameter optimization with momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

In the diffusion process, the total number of diffusion steps was set to $T = 1000$, and a linear noise scheduling strategy was adopted for Gaussian noise injection. The latent-space downsampling factor was set to $s = 8$, which effectively reduced computational complexity while preserving essential semantic and structural information.

For all experiments, the model parameters were initialized using pre-trained latent diffusion weights, and mixed-precision training was adopted to further improve training efficiency and GPU memory utilization.

B. Evaluation Metrics

To comprehensively evaluate the performance of the proposed multi-instance line art colorization framework, four widely used image quality assessment metrics are adopted, including Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

FID is employed to measure the distribution discrepancy between generated images and real images in a high-dimensional feature space. A lower FID value indicates that the generated image distribution is closer to the real image distribution, implying higher generation quality and better visual realism.

PSNR is utilized to evaluate the pixel-level similarity between the generated image and the reference image, which is defined as:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{L^2}{\mathrm{MSE}}, \tag{35}$$

where, L denotes the maximum possible pixel intensity value and MSE represents the mean squared error between the generated image and the reference image. Higher PSNR values indicate better reconstruction fidelity and lower pixel-level distortion.

SSIM is adopted to measure structural consistency between generated images and reference images in terms of luminance, contrast, and structural information. It is formulated as:

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{36}$$

where, μ_x and μ_y denote the mean intensities of two images, σ_x^2 and σ_y^2 represent the variances, σ_{xy} denotes the covariance, and C_1 and C_2 are stability constants. An SSIM value closer to 1 indicates better structural preservation and visual consistency.

LPIPS is introduced to evaluate perceptual similarity between generated images and reference images based on deep feature representations extracted from neural networks. It is defined as:

$$\mathrm{LPIPS} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}^l_{hw} - \hat{x}^l_{hw}) \right\|_2^2, \tag{37}$$

where, l denotes different network layers, $\hat{x}^l$ and $\hat{y}^l$ represent normalized feature maps, w_l denotes the learned layer weight, and H_l and W_l correspond to the spatial dimensions of the feature maps. Lower LPIPS values indicate higher perceptual similarity and better visual quality.

In summary, lower FID and LPIPS values together with higher PSNR and SSIM values indicate that the generated images are closer to the reference images in terms of distribution consistency, pixel-level reconstruction accuracy, structural preservation, and perceptual quality. These metrics collectively provide a comprehensive evaluation of the proposed method for multi-instance line art colorization tasks.

C. Comparative Experimental Results

To validate the effectiveness of the proposed framework, extensive comparisons were conducted against several representative line art colorization methods. In all comparison methods, the complete reference image was directly used as the conditional input, whereas the proposed framework utilized instance-level information extracted from the same reference image for multi-instance conditional guidance.

For GAN-based approaches, RSIC [16] and SGA [17] were selected as representative comparison methods. Both approaches employ generative adversarial networks for line art colorization, while adopting different network architectures and training strategies. In addition, for diffusion-based methods, AnimeDiffusion [18], ColorizeDiffusion [19], and MangaNinja [20] were adopted for comparison. These methods represent recent advances in diffusion-based line art colorization and provide strong baselines for evaluating color consistency, structural fidelity, and detail restoration under multi-instance conditions.

The quantitative comparison results on the Sakuga and ATD-12K datasets are summarized in Table II.

As shown in Table II, the proposed structure-aware multi-instance line art colorization framework achieves significant performance improvements on both the Sakuga and ATD-12K datasets.

Among the GAN-based approaches, RSIC and SGA obtain relatively high FID values and moderate PSNR and SSIM scores, while their LPIPS values remain comparatively large. These results indicate that GAN-based methods still suffer from color inconsistency and structural distortion, especially in multi-instance scenes or regions containing complex color distributions. In many cases, color overflow occurs around instance boundaries, leading to loss of structural details and degraded visual coherence.

Compared with GAN-based methods, diffusion-based approaches demonstrate clear improvements in image quality and structural preservation. AnimeDiffusion, ColorizeDiffusion, and MangaNinja achieve lower FID values ranging from 25.7 to 22.8 on the Sakuga dataset and from 28.5 to 25.3 on the ATD-12K dataset. Meanwhile, PSNR and SSIM values are improved to the ranges of 24.8–25.4 and 0.801–0.817, respectively, while LPIPS values decrease to 0.165–0.176. These results demonstrate the superiority of diffusion models in maintaining global color consistency and structural fidelity. However, existing diffusion-based methods still exhibit

TABLE II. COMPARATIVE EXPERIMENTAL RESULTS ON THE SAKUGA AND ATD-12K DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	Sakuga				ATD-12K			
	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RSIC	42.5	21.3	0.732	0.241	48.7	20.9	0.715	0.257
SGA	38.9	22.1	0.749	0.227	44.3	21.7	0.731	0.243
AnimeDiffusion	25.7	24.8	0.801	0.176	28.5	24.1	0.788	0.189
ColorizeDiffusion	23.4	25.2	0.813	0.169	26.1	24.6	0.798	0.181
MangaNinja	22.8	25.4	0.817	0.165	25.3	24.8	0.802	0.178
Proposed Method	**15.2**	**27.1**	**0.856**	**0.121**	**17.4**	**26.3**	**0.841**	**0.135**

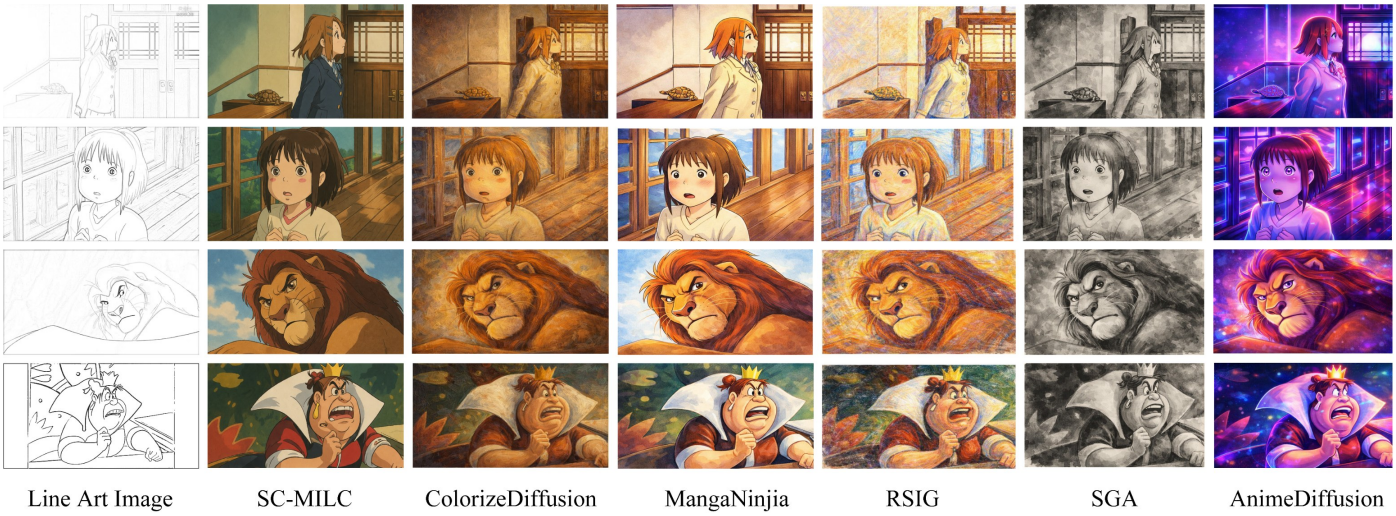


Fig. 3. Qualitative comparison results of different line art colorization methods. The proposed method achieves superior structural consistency, color fidelity, and instance-level detail restoration compared with existing GAN-based and diffusion-based approaches.

limitations in instance-level color mapping, particularly for small-scale details and complex multi-instance scenarios.

In contrast, the proposed method achieves the best performance across all evaluation metrics on both datasets. On the Sakuga dataset, the proposed framework reduces the FID score to 15.2, while improving PSNR and SSIM to 27.1 and 0.856, respectively, with LPIPS further reduced to 0.121. Similarly, on the ATD-12K dataset, the proposed framework achieves FID, PSNR, SSIM, and LPIPS values of 17.4, 26.3, 0.841, and 0.135, respectively. These results indicate that the proposed framework effectively preserves structural details while achieving more accurate instance-level color mapping.

In particular, the collaborative interaction between the latent generative backbone and the instance semantic fusion module enables the generated images to maintain high perceptual similarity to the reference images, especially in regions containing small-scale instances or complex color distributions. The proposed structure-aware guidance mechanism further improves edge sharpness and spatial consistency during the diffusion denoising process.

Furthermore, the qualitative comparison results shown in Fig. 3 provide intuitive visual evidence of the performance differences among various methods.

GAN-based methods can generate basic color regions; however, they often exhibit noticeable color overflow and inconsistent color distributions around instance boundaries and small-scale regions. For example, colors from clothing or accessories frequently bleed into adjacent regions, resulting in poor global visual consistency and insufficient detail restoration.

Diffusion-based methods achieve more natural global color distributions and better structural preservation compared with GAN-based methods. Nevertheless, they still encounter difficulties in accurately restoring local color details in complex multi-instance scenarios, and color inconsistencies among different instances remain observable in some cases.

By contrast, the proposed framework demonstrates clear visual advantages. Through the integration of the latent generative backbone, structure-aware guidance, and instance semantic fusion mechanisms, the generated images preserve fine line structures while achieving accurate instance-level color mapping. Particularly in multi-instance scenes containing complex color regions and small objects, the proposed method produces natural color distributions, sharp structural boundaries, and highly consistent color representations across instances.

Moreover, the proposed framework effectively suppresses color bleeding and feature interference among neighboring instances, thereby exhibiting stronger controllability and superior detail restoration capability in multi-instance line art colorization tasks.

Overall, both quantitative and qualitative results demonstrate that the proposed framework significantly improves instance-level controllability, structural fidelity, and overall color consistency, highlighting its effectiveness and practical

value for multi-instance line art colorization applications.

D. Ablation Experiments

To verify the effectiveness of each core component and the unified optimization objective in the proposed framework, a series of ablation experiments were conducted on the Sakuga and ATD-12K datasets. The experiments were designed as follows:

- Experiment A: The structure-aware guidance module (SGM) was removed while retaining the latent generative backbone (LGB) and the instance semantic modeling and feature fusion module (ISMFF). This experiment evaluates the contribution of structural guidance to contour preservation and edge fidelity.
- Experiment B: The ISMFF module was removed while retaining the LGB and SGM modules. In this setting, only global reference information was utilized without instance-level semantic modeling and feature fusion. This experiment evaluates the effectiveness of instance-level semantic guidance for precise multi-instance color control.
- Experiment C: The LGB module was removed while retaining SGM and ISMFF. In this case, latent-space diffusion generation was not employed, and the framework directly performed color mapping based on structural and semantic guidance. This experiment evaluates the importance of latent-space generation for overall image quality and generation stability.
- Experiment D: The complete framework was adopted, including LGB, SGM, ISMFF, and the unified optimization objective. This configuration was used to evaluate the collaborative effectiveness of all proposed modules.

As shown in Table III and Table IV, each module contributes significantly to the overall performance of the proposed framework.

On the Sakuga dataset, Experiment A removes the SGM module while retaining LGB and ISMFF. Without structural guidance, the generated images exhibit blurred edge structures and insufficient contour consistency, resulting in an FID score of 28.5, a PSNR of 24.3, an SSIM of 0.790, and an LPIPS value of 0.182. These results demonstrate that structural information plays an important role in preserving edge sharpness and structural fidelity.

Experiment B removes the ISMFF module while retaining LGB and SGM. Although the overall colorization results remain visually reasonable, obvious color deviations appear in multi-instance regions. The obtained FID, PSNR, and SSIM values indicate that instance-level semantic modeling is crucial for precise multi-instance color control and semantic consistency.

Experiment C removes the latent generative backbone and only retains SGM and ISMFF. In this setting, the overall generation capability deteriorates significantly, with the FID score increasing to 36.8 and LPIPS increasing to 0.215, while PSNR and SSIM decrease substantially. These results demonstrate that the latent diffusion generation backbone provides the fundamental generation capability and stability for the proposed framework.

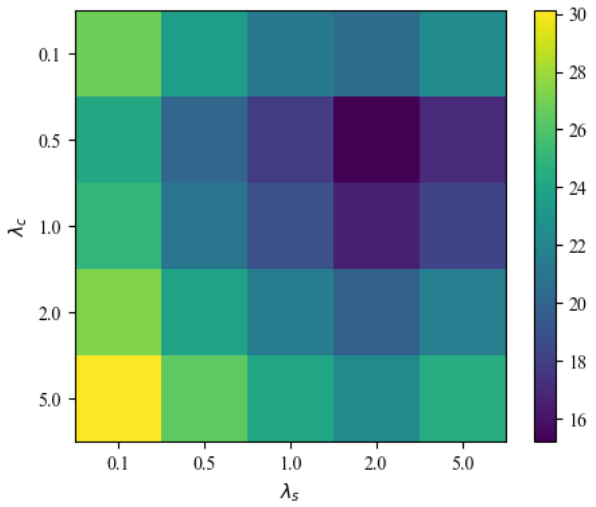


Fig. 4. FID heatmap under different λ_s and λ_c with a fixed λ_g.

Experiment D adopts the complete framework and achieves the best performance across all evaluation metrics. The generated images exhibit clear structural boundaries, accurate color mapping, and high visual consistency, demonstrating the effectiveness of the collaborative interaction among all proposed modules.

Similar trends can also be observed on the ATD-12K dataset. Removing SGM in Experiment A leads to weaker structural preservation and degraded edge quality. Removing ISMFF in Experiment B results in noticeable color inconsistencies among different instances. Meanwhile, removing LGB in Experiment C causes the most severe performance degradation in overall image quality and generation stability. By contrast, the complete framework in Experiment D achieves the best quantitative performance, further validating the effectiveness of the proposed module collaboration strategy.

Overall, the ablation experiments demonstrate that the LGB module provides stable latent-space generation capability and robust color representation, the SGM module enhances structural consistency and edge fidelity, and the ISMFF module enables accurate multi-instance semantic alignment and color control. Moreover, the unified optimization objective further improves high-frequency detail restoration and overall visual quality. These results comprehensively verify the necessity and effectiveness of each proposed component for high-quality and controllable multi-instance line art colorization.

E. Parameter Sensitivity Analysis

To further analyze the coupling relationship among different constraint terms in the proposed optimization objective, a two-dimensional grid search was conducted for λ_s and λ_c while fixing λ_g. The corresponding FID heatmap is illustrated in Fig. 4.

It can be observed that the model performance exhibits a distinct asymmetric distribution in the parameter space, and

TABLE III. ABLATION EXPERIMENTAL RESULTS ON THE SAKUGA DATASET.

Exp.	Modules			Metrics			
	LGB	SGM	ISMFF	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
A	✓		✓	28.5	24.3	0.790	0.182
B	✓	✓		25.1	25.2	0.810	0.170
C		✓	✓	36.8	22.1	0.752	0.215
D	✓	✓	✓	**15.2**	**27.1**	**0.856**	**0.121**

TABLE IV. ABLATION EXPERIMENTAL RESULTS ON THE ATD-12K DATASET.

Exp.	Modules			Metrics			
	LGB	SGM	ISMFF	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
A	✓		✓	31.1	23.8	0.775	0.196
B	✓	✓		27.6	24.7	0.795	0.182
C		✓	✓	39.5	21.7	0.739	0.231
D	✓	✓	✓	**17.4**	**26.3**	**0.841**	**0.135**

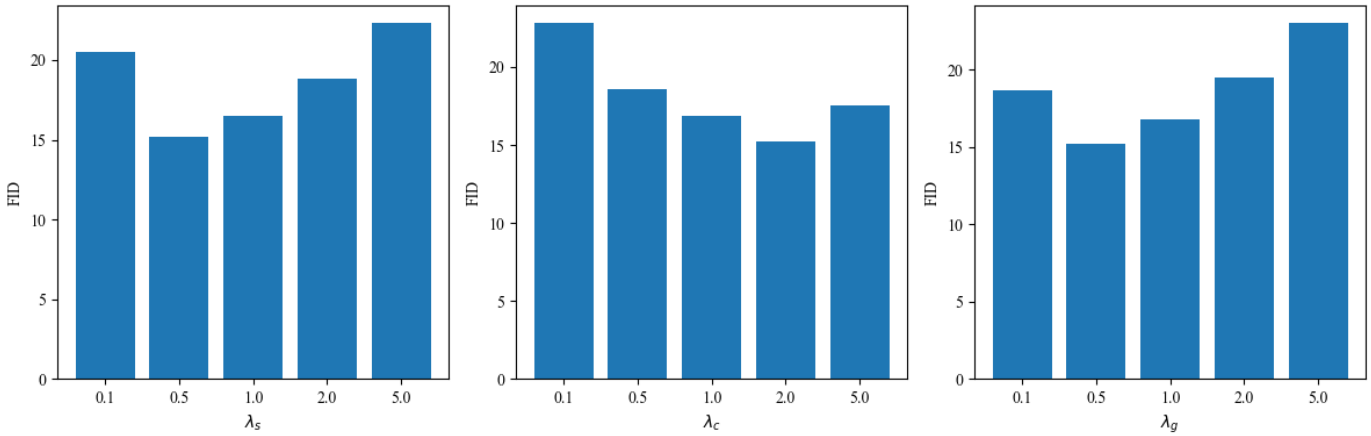


Fig. 5. Results of the univariate sensitivity analysis for different constraint weights.

the optimal performance is achieved around specific non-uniform parameter combinations rather than simple equal-weight settings. This phenomenon indicates that different constraint terms contribute unequally during the optimization process.

Specifically, moderately increasing λ_c helps improve color consistency, enabling the generated images to better match the reference instances. In contrast, λ_s should be controlled within a relatively small range to avoid excessive structural constraints that may suppress detail flexibility. When λ_s becomes overly large, the model excessively emphasizes sketch structures and weakens color representation capability. Conversely, excessively large λ_c values tend to produce overly concentrated color distributions, thereby reducing visual naturalness.

These results demonstrate that a significant trade-off exists between structural preservation and color consistency. Meanwhile, the instance fusion regularization term λ_g achieves the best performance within a moderate range, effectively balancing local instance features and global semantic representations. Overall, the three constraint terms do not contribute equally but instead cooperate through non-uniform weighting strategies, thereby validating the necessity and rationality of the proposed unified optimization objective.

To further investigate the influence of each constraint weight individually, univariate sensitivity experiments were conducted for λ_s, λ_c, and λ_g, while fixing the remaining parameters to their optimal settings. The experimental results are shown in Fig. 5.

It can be observed that all three parameters exhibit a similar trend in which the performance first improves and then deteriorates as the parameter value increases. However, the optimal ranges differ significantly among the three parameters.

The optimal value of λ_s is relatively small, indicating that moderate structural constraints are beneficial for preserving edge clarity, whereas excessive structural guidance suppresses fine-detail generation. In contrast, the optimal value of λ_c is comparatively larger, demonstrating that color consistency plays a more important role in multi-instance line art colorization tasks. Meanwhile, λ_g achieves the best performance within a moderate range. Excessively large fusion regularization tends to over-smooth semantic features and weakens local detail representation.

These results further confirm that different constraint terms play distinct roles during optimization. Therefore, the three constraints should be collaboratively adjusted through non-uniform weighting in order to achieve an effective balance

among structural fidelity, color consistency, and semantic fusion capability.

V. Conclusion and Future Work

This study proposes a structure-aware line art colorization framework based on latent diffusion to address the challenges of insufficient structural preservation, inaccurate color mapping, and semantic inconsistency in line art colorization tasks. The proposed framework introduces a structure-aware constraint mechanism to enhance contour fidelity and spatial consistency during the diffusion generation process. Meanwhile, an instance semantic modeling and feature fusion strategy is designed to improve local color consistency and instance-level controllability. Furthermore, a unified optimization objective is constructed to collaboratively optimize structural constraints, color consistency, and semantic fusion regularization within a unified latent-space generation framework.

Experimental results on publicly available datasets demonstrate that the proposed method achieves superior performance compared with existing approaches in terms of FID, PSNR, SSIM, and LPIPS metrics. The generated colorization results exhibit clearer structural details, more natural color distributions, and stronger semantic consistency. In particular, the proposed framework effectively improves color controllability and detail restoration capability in complex multi-instance scenarios.

Overall, the proposed framework provides an effective solution for high-quality and controllable line art colorization, and offers a feasible technical reference for related image generation tasks and electronic measurement vision applications.

In future work, the proposed framework can be further extended in several directions. First, introducing temporal consistency modeling may improve color stability in animation sequence generation tasks. Second, incorporating multimodal conditional guidance, such as text descriptions or style priors, could further enhance controllability and diversity in color generation. In addition, lightweight diffusion architectures and efficient inference strategies may be explored to improve computational efficiency and facilitate practical deployment in real-time intelligent visual systems.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62573298, and by the Guangdong Provincial Key Laboratory under Grant No. 2023B1212060076.

References

[1] Q. Gao, M. Wu, X. Qin, and L. Hua, "Machine vision driven magnetic particle inspection technology: principles, applications and trends," *Measurement Science and Technology*, vol. 37, no. 3, p. 032001, 2026.

[2] P. Wu, X. He, W. Dai, J. Zhou, Y. Shang, Y. Fan, and T. Hu, "A review on research and application of al-based image analysis in the field of computer vision," *IEEE Access*, 2025.

[3] Z. Li, Z. Geng, Z. Kang, W. Chen, and Y. Yang, "Eliminating gradient conflict in reference-based line-art colorization," in *European conference on computer vision*. Springer, 2022, pp. 579–596.

[4] H. Carrillo, M. Clément, A. Bugeau, and E. Simo-Serra, "Diffusart: Enhancing line art colorization with conditional diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3486–3490.

[5] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen, "Diffusion models for image restoration and enhancement: A comprehensive survey," *International Journal of Computer Vision*, vol. 133, no. 11, pp. 8078–8108, 2025.

[6] Y. Zhang, Y. Ma, B. Wang, Q. Chen, and Z. Wang, "Magiccolor: Multi-instance sketch colorization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 15 205–15 217.

[7] C. Li, "Semantically aware style-controlled animation line art colorization using conditional gans with gcn and attention mechanisms," *Informatica*, vol. 49, no. 29, 2025.

[8] Y. Cao, X. Duan, X. Meng, P. Mok, P. Li, and T.-Y. Lee, "Computer-aided colorization state-of-the-science: A survey," *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[9] J.-X. Chen, L. Lo, S.-Y. Lu, L. Zou, W.-H. Cheng, J. Huh, and S. Lee, "Seco: Semantic-guided multimodal color splash effects," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2026.

[10] J. K. Pandya, S. S. Khandelwal, R. K. Tipu, and K. S. Pandya, "Advancing water quality management: an integrated approach using ensemble machine learning and real-time interactive visualization," *IEEE Access*, 2025.

[11] M. Fuest, P. Ma, M. Gui, J. Schusterbauer, V. T. Hu, and B. Ommer, "Diffusion models and representation learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[12] W. Jiang, Y. Li, Z. Yi, M. Chen, and J. Wang, "Multi-instance imbalance semantic segmentation by instance-dependent attention and adaptive hard instance mining," *Knowledge-Based Systems*, vol. 304, p. 112554, 2024.

[13] S. Fatima, S. Ali, and H.-C. Kim, "A comprehensive review on multiple instance learning," *Electronics*, vol. 12, no. 20, p. 4323, 2023.

[14] Z. Pan, "Sakuga-42m dataset: Scaling up cartoon research," *arXiv preprint arXiv:2405.07425*, 2024.

[15] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, and Z. Liu, "Deep animation video interpolation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6587–6595.

[16] J. Lee, E. Kim, Y. Lee *et al.*, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[17] Z. Li, Z. Geng, Z. Kang *et al.*, "Eliminating gradient conflict in reference-based line-art colorization," in *European Conference on Computer Vision*. Springer, 2022.

[18] Y. Cao, X. Meng, P. Y. Mok *et al.*, "Animediffusion: Anime face line drawing colorization via diffusion models," *arXiv preprint arXiv:2303.11137*, 2023.

[19] D. Yan, L. Yuan, E. Wu *et al.*, "Colorizediffusion: Adjustable sketch colorization with reference image and text," *arXiv preprint arXiv:2401.01456*, 2024.

[20] Z. Liu, K. L. Cheng, X. Chen *et al.*, "Manganinja: Line art colorization with precise reference following," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

C-TriLoRA: Cross-Corpus Kazakh SER via Tri-Factor LoRA and CORAL

Bakdaulet Kynabay[1]*, Aimoldir Aldabergen[2], Shirali Kadyrov[3]
Faculty of Engineering and Natural Sciences, SDU University, Almaty, Kazakhstan[1]
School of Sciences and Humanities, Nazarbayev University, Almaty, Kazakhstan[2]
Department of General Education, New Uzbekistan University, Tashkent, Uzbekistan[3]

***Abstract*—Speech Emotion Recognition (SER) in low-resource languages deals with the scarcity of labeled corpora and the instability of learned representations when transferred across diverse recording conditions and speaker demographics. This study introduces Conditional Tri-Factor Low-Rank Adaptation (C-TRILORA), a multi-task architecture that jointly performs automatic speech recognition (ASR) and SER on Kazakh speech while generalizing reliably across corpora. The proposed model extends a pre-trained Whisper encoder–decoder backbone through three primary innovations: a Tri-LoRA routing module that disentangles lexical, emotional, and speaker latent factors; CORAL domain alignment that matches second-order statistics between source and target domains without target labels; and a gradient reversal layer (GRL) that suppresses speaker-identity information. Experimental evaluations on the KazEmoTTS and ENU KEMO datasets demonstrate that C-TRILORA achieves a competitive in-domain Macro-F1 of 86.09% and significantly outperforms standard baselines in cross-corpus conditions (41.44% Macro-F1 versus 37.34% for the Shared-Head baseline). McNemar and Wilcoxon signed-rank tests confirm that explicit factor disentanglement is essential for cross-corpus robustness. These results show that separating speech components effectively mitigates negative transfer, making C-TRILORA a practical approach for low-resource SER deployment.**

***Keywords*—Speech emotion recognition; low-resource languages; parameter-efficient fine-tuning; domain adaptation; multi-task learning; disentangled representations**

I. Introduction

Over the past decade, deep learning has transformed speech emotion recognition, yet most of this progress has concentrated on a handful of well-resourced languages such as English and Mandarin [1], [2]. Languages with limited data infrastructure remain largely overlooked. Kazakh is one such case: a Turkic language with around 18 million speakers whose emotional speech has received little systematic attention, partly because publicly available, labeled corpora for it are scarce [3]. Without adequate training data, building SER systems that work reliably in Kazakh is difficult from the outset.

Beyond data scarcity, a pervasive challenge in deploying SER systems is corpus dependency, commonly referred to as the cross-corpus generalization problem. Acoustic representations learned by deep neural networks are notoriously brittle; they tend to conflate paralinguistic affective cues with domain-specific artifacts such as recording conditions, background noise, and speaker demographics [4], [5]. When a model trained on a specific source dataset is evaluated on an unseen target dataset, performance typically degrades significantly due to this covariate shift. Existing approaches to mitigate domain shift, such as adversarial domain adaptation and self-supervised learning, often treat the acoustic signal as a monolithic entity [6], [7], [8]. They generally fail to explicitly disentangle the underlying factors of speech, namely, lexical content, emotional prosody, and speaker identity, which limits their effectiveness in highly variable, cross-corpus scenarios.

To address these limitations, recent advancements have increasingly focused on mitigating the computational costs of large pre-trained models and the performance degradation caused by domain shifts in data-scarce environments. For instance, parameter-efficient fine-tuning (PEFT) techniques, particularly Low-Rank Adaptation (LoRA), have been successfully applied to foundation models like Whisper to extract robust emotional representations with minimal trainable parameters [9]. Furthermore, researchers have recently leveraged transfer learning and transformer-based paradigms to enhance cross-lingual and cross-corpus generalization [10]. However, aligning the feature distributions between mismatched source and target datasets remains a critical bottleneck. Consequently, contemporary cross-corpus SER frameworks are increasingly adopting advanced domain adaptation strategies, such as adversarial domain generalization [11] and classification inconsistency alignment [12]. These ongoing trajectories underscore the necessity of synergizing PEFT with unsupervised domain alignment and explicit factor disentanglement.

Within speech emotion recognition specifically, self-supervised transformer representations combined with lightweight fine-tuning have improved robustness in data-scarce conditions [13], and multi-task learning coupled with subdomain adaptation has been used to bridge the feature-distribution gap between mismatched corpora [14].

Motivated by these findings, this study develops C-TRILORA, a multi-task architecture for cross-corpus Kazakh SER. Rather than applying standard fine-tuning to a pre-trained Whisper backbone [15], we introduce a Tri-LoRA routing module that splits the shared speech representation into three separate streams:lexical, emotional, and speaker; each processed by its own low-rank adapter. This separation is the core design choice, aimed directly at the corpus dependency problem. The three main contributions of the work are:

- Architectural innovation: A Tri-LoRA routing module is integrated into a pre-trained Whisper backbone, explicitly decomposing shared representations into disentangled lexical, emotional, and speaker latent

*Corresponding author

factors.

- Unsupervised domain alignment: Correlation Alignment (CORAL) is integrated directly into the emotional latent space, successfully matching second-order statistics between source (KazEmoTTS) and target (ENU KEMO) domains without requiring labeled target data [3], [15]. This is coupled with a gradient reversal layer (GRL) to suppress speaker-identity leakage.
- Validation in a low-resource context: The efficacy of this approach is demonstrated on Kazakh speech, showing that explicit factor disentanglement significantly mitigates negative transfer and outperforms standard baseline models in cross-corpus SER tasks.

II. Related Work

A. Speech Emotion Recognition in Low-Resource Languages

Modern SER systems achieve strong results on richly resourced languages, supported by large corpora and by two decades of feature-engineering and benchmark work [1], [2]. For low-resourced languages, including Kazakh, progress is constrained by the absence of large labeled emotional corpora. The KazEmoTTS corpus [3] was released primarily for emotional text-to-speech rather than recognition, and recent multimodal efforts target low-resource settings explicitly [20]. A complementary direction reduces the labelled-data requirement through transfer learning and transformer architectures [10], [8]. These studies establish that low-resource SER is feasible, but remains sensitive to the conditions under which the training data were collected.

B. Cross-Corpus SER and Unsupervised Domain Adaptation

Cross-corpus SER addresses the realistic case in which training and test data are taken from different corpora with mismatched recording environments, speakers, languages, and elicitation styles. Classical solutions align feature distributions through subspace learning or transfer factorization, while deep approaches minimise an explicit domain discrepancy. Domain-adversarial training (DANN) [7] learns domain-invariant features through a gradient reversal layer, and Deep CORAL [16] aligns the second-order statistics of source and target representations. Self-supervised adversarial domain adaptation has further been applied jointly to cross-corpus and cross-language SER [25]. Recent transformer-based methods push this further: adversarial domain-generalised transformers [11], classification inconsistency alignment networks [12], multitask transformers that jointly model auxiliary tasks for cross-corpus transfer [22], source-free contrastive adaptation networks [27], and the mining of corpus-invariant emotional acoustic features [26]. A separate line of work disentangles speaker identity from affect through adversarial speaker-invariant representation learning [28], which directly motivates the speaker-suppression component used in this work. A recurring limitation of most of these methods is that they treat the acoustic embedding as a single monolithic vector and align it as a whole, without separating the lexical, affective, and speaker components that respond differently to domain shift.

C. Parameter-Efficient Fine-Tuning of Speech Foundation Models

Large self-supervised and weakly-supervised speech models such as wav2vec 2.0 [6], HuBERT [24], WavLM [23], and Whisper [17] provide powerful general-purpose representations, but full fine-tuning is expensive and prone to catastrophic forgetting in data-scarce regimes. Compact task-adapted variants such as Vesper [29], and parameter-efficient fine-tuning: Low-Rank Adaptation (LoRA) [18] in particular – instead update a small number of parameters while freezing the backbone, and have been shown effective for SER [9]. Existing PEFT-for-SER work, however, adapts the backbone with a single shared adapter and does not couple PEFT with explicit factor disentanglement or unsupervised cross-corpus alignment.

D. Research Gap

In summary, prior cross-corpus SER methods align a monolithic embedding, PEFT-SER methods use a single shared adapter without domain alignment, and low-resource SER work rarely evaluates cross-corpus transfer at all. The present work closes this gap by combining, within a single parameter-efficient model, three elements:

- A Tri-LoRA router that explicitly disentangles lexical, emotional, and speaker factors;
- CORAL alignment applied selectively to the emotional factor;
- Adversarial speaker suppression.

The resulting combination is validated on a genuinely low-resource language (Kazakh) under cross-corpus evaluation.

III. Methods

A. Problem Formulation

The fundamental objective of this study is to learn a robust emotion classifier capable of generalizing across different speech corpora without requiring target-domain labels during training. Let Eq. (1) denote the source corpus, where x_i represents a raw speech waveform, $y_i^e \in \{0, \ldots, E-1\}$ is the emotion label, y_i^t is the word-level transcript, and y_i^s is the speaker identity. Concurrently, let Eq. (2) denote the unlabelled target corpus utilized exclusively for domain alignment. The goal is to optimize a classification function, Eq. (3), that maintains high predictive accuracy on both the source domain D_S and the unseen test partition of the target domain D_T.

$$D_S = \{(x_i, y_i^e, y_i^t, y_i^s)\}_{i=1}^{N} \tag{1}$$

$$D_T = \{x_j\}_{j=1}^{M} \tag{2}$$

$$f : x \longmapsto y^e \tag{3}$$

B. Backbone Architecture: Kazakh Whisper

The proposed Conditional Tri-Factor Low-Rank Adaptation (C-TRILORA) architecture, as shown in Fig. 1, utilizes a pre-trained encoder–decoder backbone based on the Whisper architecture, which was originally developed as a robust, large-scale multilingual speech recognition system trained on weakly

supervised data [17]. Specifically for this study, the model is initialized using a Kazakh-adapted Whisper-small checkpoint [15], which was previously fine-tuned on 335 hours of Kazakh read speech. The encoder module, denoted as $\mathcal{E}$, maps an input log-Mel spectrogram $X \in \mathbb{R}^{80\times T}$ into a sequence of hidden states $H = \mathcal{E}(X) \in \mathbb{R}^{T'\times d}$, where the hidden dimension $d = 768$. To interface with the subsequent routing modules, a mean-pooled representation $\bar{h}$ is computed across the temporal dimension, as in Eq. (4):

$$\bar{h} = \frac{1}{T'}\sum_{t=1}^{T'} H_t \tag{4}$$

Whisper was chosen over encoder-only self-supervised models such as wav2vec 2.0 [6], HuBERT [24], and WavLM [23] for three reasons. First, it is an encoder–decoder model, so it natively supports the auxiliary ASR objective that grounds the lexical factor; encoder-only backbones would require an externally attached decoder for sequence-to-sequence transcription. Second, Whisper is trained on large-scale weakly-supervised, multi-condition data [17], which yields representations that are comparatively robust to the channel and noise variation that characterises cross-corpus evaluation. Third, and decisively for the low-resource setting, a Kazakh-adapted Whisper-small checkpoint fine-tuned on 335 hours of Kazakh speech is publicly available [15]; no comparable in-language checkpoint exists for the alternative backbones, so initialising from Whisper provides a substantially stronger starting point for Kazakh.

C. Parameter-Efficient Fine-Tuning and Tri-LoRA Routing

To adapt the Whisper backbone efficiently without catastrophic forgetting or excessive computational overhead, Low-Rank Adaptation (LoRA) is applied to all attention projection matrices (W_Q, W_K, W_V, W_O) in both the encoder and the decoder [18]. For any given frozen pre-trained weight matrix W_0, the adapted weight W' is computed as in Eq. (5):

$$W' = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \tag{5}$$

where, $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ are trainable low-rank matrices, $r = 16$ is the rank, and $\alpha = 32$ is a scaling factor. This parameter-efficient strategy reduces the trainable parameter count to approximately 0.31% of the total backbone. The routing module and the three auxiliary heads add only a small number of further trainable parameters on top of the LoRA adapters; a detailed profiling of the additional training-time and inference overhead they introduce is identified as a direction for future work.

The core architectural innovation of this framework is the Tri-LoRA Routing Module. Instead of relying on a single shared embedding, this module decomposes the pooled representation $\bar{h}$ into three semantically specialized, disentangled latent factors: the lexical factor z_{lex}, the emotional factor z_{emo}, and the speaker factor z_{spk}. Each factor is processed through an independent low-rank adapter stream gated by learned soft routing weights, ensuring that the model explicitly isolates affective cues from confounding linguistic and demographic variables.

D. Unsupervised Domain Alignment and Optimization

To directly mitigate corpus dependency and domain shift, Correlation Alignment (CORAL) is integrated into the emotional latent space [16]. The CORAL objective minimizes the Frobenius-norm distance between the covariance matrices (second-order statistics) of the source and target z_{emo} distributions. Furthermore, to suppress the leakage of speaker-identity information into the emotion representations, a Gradient Reversal Layer (GRL) is applied to z_{emo} during adversarial training [7].

The total training objective $\mathcal{L}$ is formulated as a weighted sum of five distinct loss components:

- Automatic Speech Recognition Loss ($\mathcal{L}_{\text{ASR}}$): Cross-entropy loss from the Whisper decoder to ground z_{lex} in lexical semantics.
- Speech Emotion Recognition Loss ($\mathcal{L}_{\text{SER}}$): Weighted cross-entropy for emotion classification.
- Speaker Classification Loss ($\mathcal{L}_{\text{SPK}}$): Cross-entropy to explicitly capture speaker traits in z_{spk}.
- Counterfactual Loss ($\mathcal{L}_{\text{CF}}$): A margin-based penalty enforcing the separation of lexical and emotional variations for in-batch pairs sharing transcripts but differing in emotion.
- CORAL Loss ($\mathcal{L}_{\text{CORAL}}$): The unsupervised covariance alignment term.

After training, post-hoc temperature scaling is applied to the output logits to optimize probabilistic calibration and reduce the Expected Calibration Error (ECE) [19].

E. Speech Emotion Datasets

To evaluate the proposed C-TRILORA architecture, particularly its cross-corpus generalization capabilities, this study utilizes two distinct Kazakh speech emotion corpora:

- KazEmoTTS (Source Domain): This dataset serves as the primary labeled source domain (D_S) for training. It consists of high-quality, studio-recorded Kazakh emotional speech originally developed for text-to-speech applications [3], and comprises six emotion categories (neutral, angry, happy, sad, fear/scared, and surprise).
- ENU KEMO (Target Domain): This dataset serves as the unlabeled target domain (D_T) for the unsupervised Correlation Alignment (CORAL) and is subsequently used to evaluate cross-corpus performance. ENU KEMO features different acoustic conditions and speaker demographics compared to KazEmoTTS, providing a robust testbed for covariate shift [20].

F. Experimental Setup and Evaluation Metrics

All acoustic inputs were standardized. Raw audio waveforms were resampled to 16 kHz, and 80-dimensional log-Mel spectrograms were extracted using a 25 ms window and a 10 ms hop size to match the input requirements of the Whisper encoder.

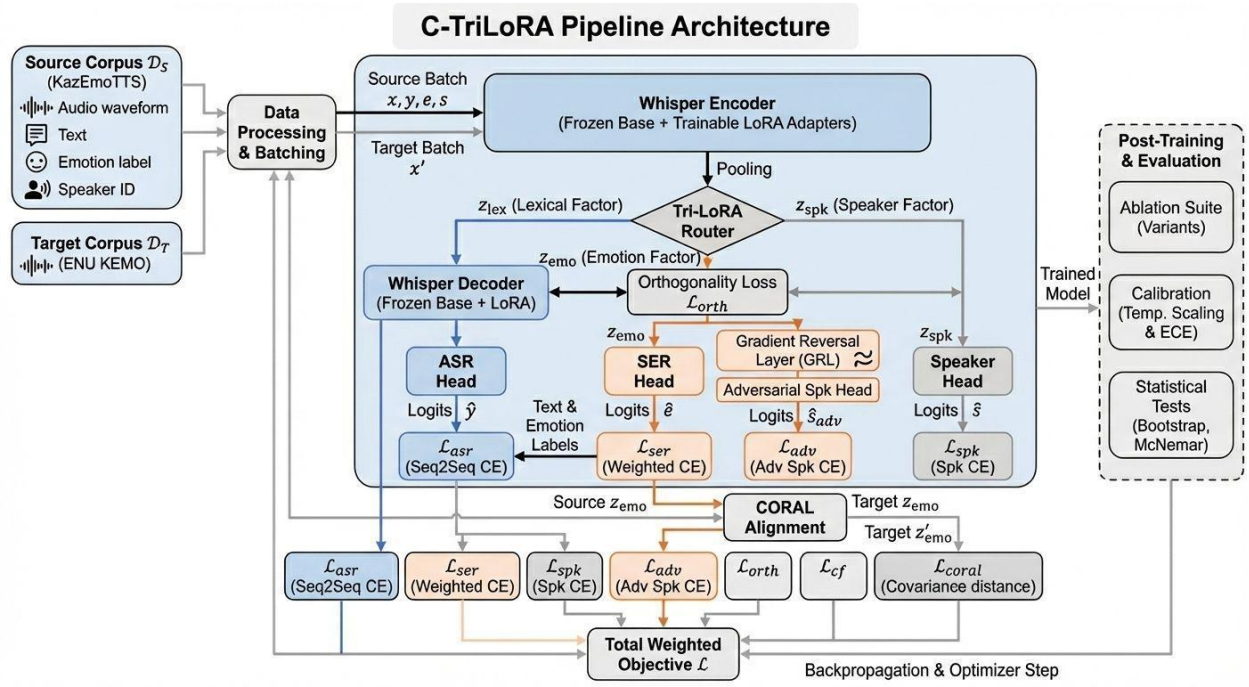


Fig. 1. The proposed C-TRILORA architecture illustrating the Tri-LoRA routing module for factor disentanglement and the CORAL module for domain alignment.

To rigorously assess the performance of the emotion classification, the primary evaluation metric used is Unweighted Average Recall (UAR). UAR is standard in speech emotion recognition tasks because it prevents majority-class bias in imbalanced datasets by calculating the accuracy independently for each emotion class and then averaging the results. Macro-F1 is reported alongside UAR throughout, as it likewise weights all emotion classes equally while being sensitive to false positives.

Additionally, to validate the statistical significance of the performance improvements yielded by the Tri-LoRA routing and CORAL components, two statistical tests were employed: *McNemar's test*, used to assess the statistical significance of differences in classification outcomes between the baseline and proposed models; and the *Wilcoxon signed-rank test*, applied to compare the paired differences in model performance across different experimental folds [21].

All models were trained using the AdamW optimizer with a learning rate of 1×10^{-4}, linear warmup over the first 10% of steps, and weight decay of 0.01, for 5 epochs with a batch size of 16 on a single NVIDIA A100 GPU (40 GB) using mixed-precision (fp16). The LoRA rank was $r = 16$ with scaling factor $\alpha = 32$. Multi-task loss weights were set empirically: $\lambda_{\text{ASR}} = 0.3$, $\lambda_{\text{SER}} = 1.0$, $\lambda_{\text{SPK}} = 0.5$, $\lambda_{\text{CF}} = 0.2$, and $\lambda_{\text{CORAL}} = 0.1$. KazEmoTTS was split 80/10/10 for training, validation, and testing; ENU KEMO was used exclusively for unsupervised domain alignment and cross-corpus evaluation.

IV. Results and Discussion

A. In-Domain and Cross-Corpus Classification Performance

The primary objective of this study was to evaluate the efficacy of the proposed C-TRILORA architecture under cross-corpus evaluation scenarios, where the model is tested on unseen target data (ENU KEMO) after being trained exclusively on the source domain (KazEmoTTS). Table I presents the classification performance across all model variants. Furthermore, the overall training stability and the convergence of the distinct multi-task loss components over the five epochs are illustrated in Fig. 2.

On the in-domain KazEmoTTS test set, the C-TRILORA model achieved strong performance with a Macro-F1 score of 86.09%. While the standard Shared Head baseline slightly outperformed the proposed model in-domain (Macro-F1 $=$ 89.69%), this is a well-documented phenomenon in representation learning. The Shared Head overfits to the joint statistics of the source domain's recording conditions and actor pool. Consequently, its representations are optimal for in-domain discrimination but highly fragile under distributional shift. Conversely, in the cross-corpus evaluation on ENU KEMO, C-TRILORA demonstrated superior robustness, achieving a Macro-F1 of 41.44% and significantly outperforming the Shared Head baseline (37.34%).

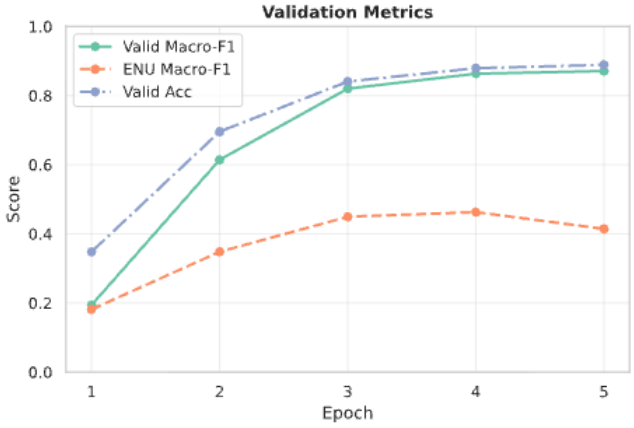
This confirms that explicitly disentangling the lexical, emotional, and speaker factors sacrifices a marginal amount of in-domain specificity in exchange for a substantial increase in cross-domain transferability. This preservation of class separability and domain-invariant clustering under covariate shift is visually corroborated by the t-SNE projections of the emotional latent space (z_{emo}) presented in Fig. 3.

B. Ablation Study and the Role of Domain Alignment

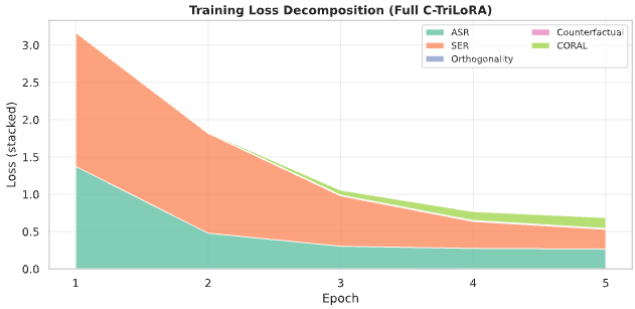
To isolate the contribution of each architectural component, an extensive ablation study was conducted. As illustrated in Fig. 4, the removal of any single component resulted

TABLE I. MACRO-F1 CLASSIFICATION PERFORMANCE OF THE C-TRILORA ARCHITECTURE AND ITS ABLATION VARIANTS.

Model Configuration	KazEmoTTS (In-Domain)	ENU KEMO (Cross-Corpus)
Shared Head (Baseline)	0.897	0.374
Full C-TRILORA (Proposed)	0.861	0.414
SER Only	0.722	0.255
No Orthogonality	0.638	0.264
No CORAL	0.615	0.211
No GRL	0.606	0.256
No Counterfactual	0.603	0.264
CORAL Only	0.578	0.239
ASR Only	0.057	0.051



(a) Validation metrics (accuracy and Macro-F1) over 5 epochs on the in-domain and cross-corpus sets.

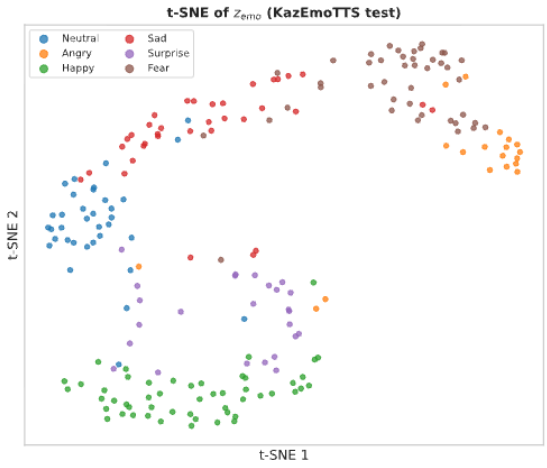


(b) Stacked decomposition of the multi-task training loss (ASR, SER, Orthogonality, Counterfactual, CORAL).

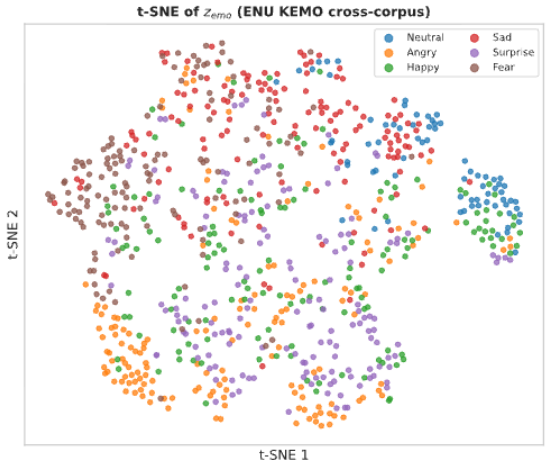
Fig. 2. Training and validation progression of the C-TRILORA architecture over 5 epochs.

in a degradation of cross-corpus performance. To rigorously validate these performance drops, statistical significance was assessed using McNemar's test. As shown in the heatmap in Fig. 5, the cross-corpus performance degradation observed in all ablation variants is statistically significant ($p < 0.05$) when compared to the full C-TRILORA architecture.

Most notably, as visualized in the waterfall chart in Fig. 6, the removal of the Correlation Alignment (CORAL) objective caused a dramatic 17.5 to 19.3 percentage point absolute drop



(a) In-domain (KazEmoTTS test)



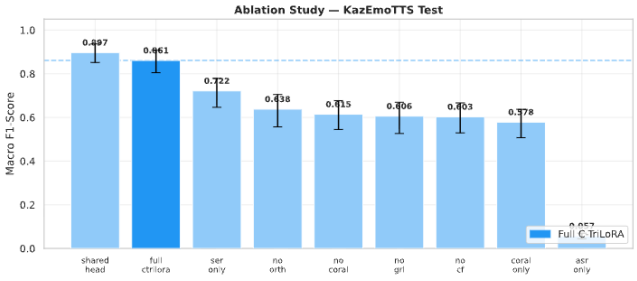
(b) Cross-corpus (ENU KEMO)

Fig. 3. t-SNE visualizations of the emotional latent space z_{emo}, illustrating the preservation of class separability under domain shift.

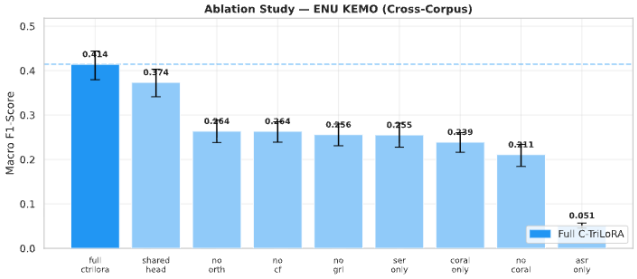
in Macro-F1 on the ENU KEMO dataset depending on the specific configuration. This finding underscores the centrality of second-order distribution matching for cross-corpus Speech Emotion Recognition (SER). This empirical result strongly aligns with established domain adaptation theory, which posits that minimizing the discrepancy in the covariance structure of source and target distributions provides a mathematical upper bound reduction on target-domain generalization error [4].

Furthermore, the CORAL-only variant only achieved a 23.91% Macro-F1. This indicates that unsupervised domain alignment alone is insufficient. It must be synergistically paired with explicit factor disentanglement to prevent negative transfer. We further note that two of these ablation configurations double as implicit comparisons to standard domain-adaptation baselines: the CORAL-only variant instantiates Deep CORAL [16], and the gradient-reversal component corresponds to domain-adversarial training (DANN) [7]. The full C-TRILORA surpasses both, indicating that explicit factor disentanglement adds value beyond these established alignment strategies; a broader empirical comparison against further

published cross-corpus systems [11], [12], [22], [27], [26] re-implemented under an identical Kazakh protocol is identified as future work.



(a) In-domain KazEmoTTS test set.



(b) Cross-corpus ENU KEMO dataset.

Fig. 4. Ablation study: Macro-F1 across model configurations.

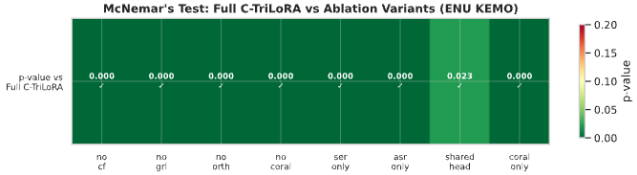


Fig. 5. Heatmap of p-values from McNemar's test, confirming the statistical significance of performance differences between the full C-TRILORA model and its ablation variants on the ENU KEMO dataset.

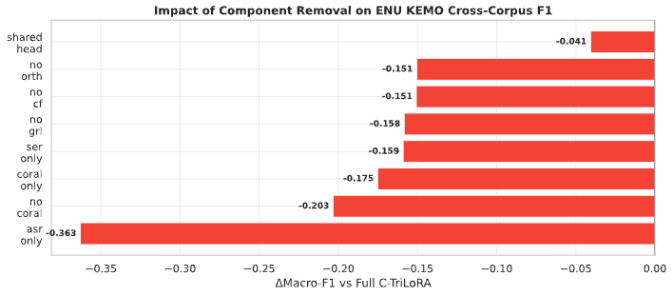
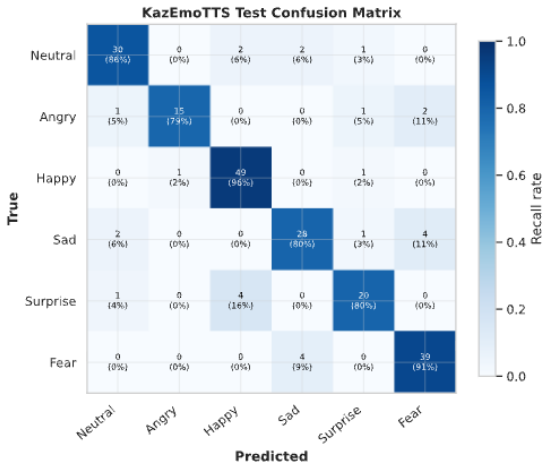


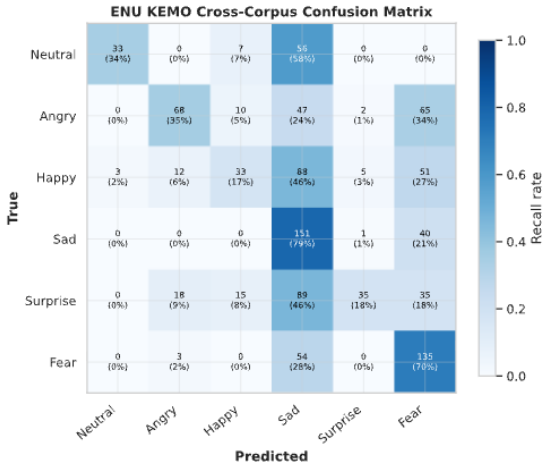
Fig. 6. Waterfall chart illustrating the absolute reduction (delta) in cross-corpus Macro-F1 on the ENU KEMO dataset when individual architectural components are removed from the full C-TRILORA model.

studies on the "valence bottleneck" in SER [5]. The confusion matrices in Fig. 7 demonstrate that high-arousal emotions with unambiguous, biologically driven prosodic signatures such as fear and neutral affect generalized relatively well across the two Kazakh corpora.

However, as highlighted by the per-class distribution in the radar plot (Fig. 8), mid-valence emotions like happiness and sadness suffered the most severe performance drops. The model exhibited systematic over-prediction of sadness and confusion between happy and sad states when tested on ENU KEMO. This aligns with broader SER literature indicating that the acoustic realizations of mid-valence emotions vary substantially across different speaker demographics, recording setups, and elicitation styles (e.g., read speech versus acted speech). Explicitly mapping these nuanced valence differences without target-domain labels remains a fundamental challenge.



(a) KazEmoTTS test set.



(b) Cross-corpus ENU KEMO.

Fig. 7. Confusion matrices of the proposed C-TRILORA model.

C. Error Analysis and Valence-Axis Shift

A detailed per-class analysis of the cross-corpus predictions reveals a systematic pattern consistent with previous theoretical

D. Probabilistic Calibration

Finally, the probabilistic calibration of the network was assessed using the Expected Calibration Error (ECE). As shown

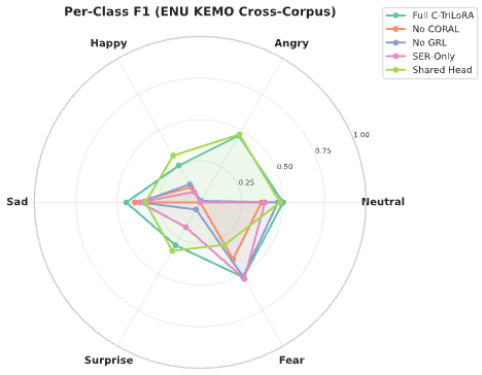
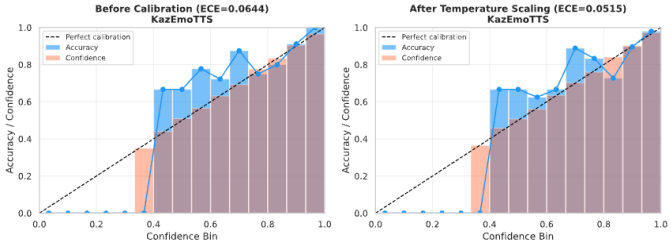


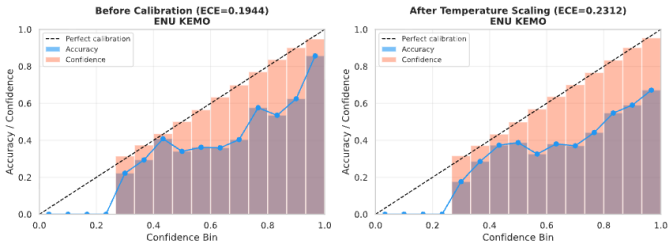
Fig. 8. Radar plot of per-class F1-scores for the ENU KEMO cross-corpus evaluation across ablation variants, demonstrating the "valence bottleneck" for mid-arousal emotions such as happiness and sadness.

in Fig. 9, before calibration, the model exhibited slight overconfidence in-domain (ECE $= 6.44\%$). Post-hoc temperature scaling successfully reduced this to 5.15%. However, under cross-corpus domain shift, the raw ECE rose substantially to 19.44%, reflecting systematic overconfidence on out-of-distribution data.

Crucially, applying the optimal in-domain temperature parameter actually worsened the cross-corpus ECE to 23.12%, confirming that standard temperature scaling does not seamlessly transfer across domains. This highlights a critical area for future research in domain-adaptive calibration for affective computing architectures.



(a) In-domain KazEmoTTS test set.



(b) Cross-corpus ENU KEMO dataset.

Fig. 9. Reliability diagrams comparing model accuracy against confidence (ECE) before and after post-hoc temperature scaling.

V. Conclusion

This study successfully addressed the challenge of cross-corpus covariate shift in low-resource speech emotion recognition through the development of the C-TRILORA architecture for Kazakh speech. The integration of a Tri-LoRA routing module separated lexical, emotional, and speaker representations, which, when combined with unsupervised CORAL domain alignment, reduced the fragility of standard acoustic models. Experimental results confirmed that this parameter-efficient approach significantly outperforms baseline models on unseen target domains by preserving domain-invariant affective features. Taken together, the results support the value of both feature disentanglement and second-order distribution matching for deploying reliable affective computing systems in low-data settings.

This study nevertheless has several limitations, which in turn define clear directions for future work. First, the evaluation is currently restricted to two Kazakh corpora; extending it to additional Kazakh datasets and, importantly, to other low-resource languages is a priority for establishing broader generality. Because the Tri-LoRA router, the CORAL alignment, and the gradient reversal layer make no language-specific assumptions (only the Whisper backbone checkpoint is language-dependent, and multilingual Whisper checkpoints are readily available), the architecture is expected to transfer to other languages, and a cross-lingual evaluation is planned. Second, beyond the internal Shared-Head baseline and the ablation variants two of which, the CORAL-only and gradient-reversal configurations, already correspond to the established Deep CORAL and DANN baselines - a comprehensive empirical comparison against the recent published cross-corpus systems surveyed in Section II, re-implemented under an identical Kazakh protocol, is an important next step. Third, the multi-task loss weights were set empirically; a systematic sensitivity analysis and joint optimisation of these weights remain to be carried out. Fourth, a detailed profiling of the computational cost: the additional training time, FLOPs, peak memory, and inference latency introduced by the routing module relative to the baseline is left for future work. Finally, the cross-corpus gains, while statistically significant, are not large in absolute terms, and the valence bottleneck for mid-arousal emotions together with domain-adaptive probabilistic calibration remain open problems for affective computing in low-data settings.

Acknowledgment

This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the project AP23487777. The authors declare that there is no conflict of interest regarding the publication of this article and confirm that the study is free of plagiarism.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used Gemini to assist with language translation, structural formatting, and proofreading to align with journal guidelines. After using this tool, the authors closely reviewed and edited the generated text to ensure scientific accuracy and take full responsibility for the final content.

References

[1] M. B. Akcay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] A. Abilbekov, S. Mussakhojayeva, R. Yeshpanov, and H. A. Varol, "KazEmoTTS: A dataset for Kazakh emotional text-to-speech synthesis," in *Proc. LREC-COLING*, Torino, Italy, 2024, pp. 9626–9632.

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1–2, pp. 151–175, 2010.

[5] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Interspeech*, 2018, pp. 951–955.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.

[7] Y. Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[8] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, 2023.

[9] T. Feng and S. Narayanan, "PEFT-SER: On the use of parameter-efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *Proc. ACII*, 2023, pp. 1–8.

[10] J. Ye, Y. Wei, X.-C. Wen, C. Ma, Z. Huang, K. Liu, and H. Shan, "Emo-DNA: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition," in *Proc. 31st ACM Int. Conf. Multimedia (MM)*, 2023, doi: 10.1145/3581783.3611704.

[11] Y. Gao, L. Wang, J. Liu, J. Dang, and S. Okada, "Adversarial domain generalized transformer for cross-corpus speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 15, no. 2, pp. 697–708, 2024.

[12] X. Zhou, J. Li, Q. Yu, and Q. Wu, "Classification inconsistency alignment network for cross-corpus speech emotion recognition," in *Proc. IEEE ICASSP*, 2025, pp. 1–5.

[13] S. Kakouros, T. Stafylakis, L. Mošner, and L. Burget, "Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[14] H. Fu, Z. Zhuang, Y. Wang, C. Huang, and W. Duan, "Cross-corpus speech emotion recognition based on multi-task learning and subdomain adaptation," *Entropy*, vol. 25, no. 1, p. 124, 2023.

[15] B. Kynabay, A. Aldabergen, S. Kadyrov, and A. Shalkarbay-uly, "Fine-tuning OpenAI's Whisper model for Kazakh speech recognition," 2025. [Online]. Available: https://doi.org/10.13140/RG.2.2.29371.07205

[16] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. European Conf. Computer Vision Workshops*, 2016, pp. 443–450.

[17] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[18] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th ICML*, 2017, pp. 1321–1330.

[20] M. Altaibek, A. Zulkhazhav, B. Yergesh, G. Bekmanova, and T. Aibol, "A multimodal framework for speech emotion recognition in low-resource languages," *Journal of Artificial Intelligence and Technology*, vol. 5, pp. 354–364, 2025.

[21] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[22] C.-S. Ahn, R. Rana, C. Busso, and J. C. Rajapakse, "Multitask transformer for cross-corpus speech emotion recognition," *IEEE Trans. Affective Computing*, 2025.

[23] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[25] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self-supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affective Computing*, 2022, doi: 10.1109/TAFFC.2022.3167013.

[26] H. Lian, C. Lu, Y. Zhao, S. Li, T. Qi, and Y. Zong, "Exploring corpus-invariant emotional acoustic feature for cross-corpus speech emotion recognition," *Expert Systems with Applications*, vol. 258, p. 125162, 2024.

[27] Y. Zhao, J. Wang, C. Lu, S. Li, B. W. Schuller, Y. Zong, and W. Zheng, "Emotion-aware contrastive adaptation network for source-free cross-corpus speech emotion recognition," in *Proc. IEEE ICASSP*, 2024, pp. 11846–11850.

[28] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *Proc. IEEE ICASSP*, 2020, pp. 7144–7148.

[29] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 15, no. 3, pp. 1711–1724, 2024.

Query Recovery Attack Based on Multi-Source Leakage and Semantic Embedding in Searchable Symmetric Encryption

Xiaogang Yuan, Xinle Yang, Dezhi An
School of Cyber Security, Gansu University of Political Science and Law, Lanzhou 730070, China

***Abstract*—With the rapid development of cloud computing and big data technologies, searchable encryption has become a research hotspot. To improve search efficiency, searchable encryption algorithms may leak redundant information, allowing adversaries to launch query recovery attacks by exploiting pattern leakage in searchable symmetric encryption and infer the underlying keywords of user queries. Existing query recovery attack methods only perform query recovery under multi-source leakage patterns, without considering the semantic level; they simulate user queries with plaintext rather than ciphertext and rarely take frequency as auxiliary leakage for attacks. Meanwhile, the weights between volume and frequency mostly rely on manual configuration, without dynamic allocation of the weights for volume and frequency. Therefore, this study proposes a novel attack method named refine_atk. First, a multi-layer perceptron is used to learn the weights of volume and frequency to accurately identify and recover distinctive queries. Next, co-occurrence information is employed to correct the queries recovered in the previous step. Finally, a cost matrix is constructed using the weighted co-occurrence matrix and the semantic embedding matrix obtained by the pre-trained language model MiniLM-L6, and the remaining queries are recovered in one pass via greedy graph matching. The proposed attack achieves an attack accuracy of 95% on the Enron and Lucene datasets. The attack performance remains robust even after removing part of the similar data. The attack execution efficiency of the proposed method is significantly superior to that of the traditional schemes, yielding better performance when attacking concrete searchable symmetric encryption schemes or evaluating their security. This work provides a reference for the security evaluation and defense mechanism design of searchable symmetric encryption.**

***Keywords*—Searchable symmetric encryption; query recovery attack; pattern leakage; multi-layer perceptron (MLP)**

I. Introduction

Searchable symmetric encryption (SSE) query recovery attacks mainly utilize the queries observed by the server and the returned results to infer the keywords of the user query. It is divided into active attack and passive attack. Active attack [1], is also known as injection attack. Suppose an attacker can upload a small number of malicious files, which will leak the content of encrypted queries, allowing the attacker to obtain the client query privacy. The file injection attack proposed by Zhang Y et al. [1] only requires the injection of a small amount of files. Even in the face of the SE scheme with less leakage, the query content of the client can be completely inferred. The attack proposed by Poddar et al. [2] can be recovered with just one query issued by the user. The attack exploits file injection and automatic query replay, combined with volume leakage, to recover user queries. Zhang X et al. [3] proposed two binary-based volume injection attacks, BVA and BVMA, which reduce the number of injected files by using known keywords and infer the query content by observing the query volume. Zhang L et al. [4] proposed a volume injection attack with the optimal injection size. By using the response length and size patterns, only O(logn) injection files are required, and the total size is O(nlogn).

Passive attacks can be classified into known data attacks and similar data attacks. A known data attack refers to an attacker who has full or partial knowledge of the documents stored on the server. This study focuses on the SSE solution for single-keyword query. Cash et al. [5] proposed a counting attack, which requires the number of identical files matched by any two keywords in a known document to recover the query based on the response volume (length of the access pattern). The IKK attack [6] uses access pattern leakage to recover user queries. It assumes that the attacker has some known queries and knows the probability that any two keywords appear in the same known document. The core principle is the simulated annealing algorithm. Blackstone et al. [7] proposed a subgraph attack that exploits the document volume pattern (the number of bytes each document occupies) to recover queries. The attack proposed by Ning et al. [8] simultaneously recovers user queries and databases.

Similar data attack refers to the situation where the attacker has documents with keyword distributions similar to the user real data, and the documents in the similar data do not need to be indexed. For example, public threat intelligence datasets [9] and knowledge graphs [10] can serve as sources of similar data. The attack proposed by Liu et al. [11] only relies on the leakage of search patterns. By comparing the search frequency of each keyword on known documents with the search frequency of user queries on real documents, the underlying keywords are recovered. Oya et al. [12] pointed out that by simultaneously utilizing the access pattern and the search pattern and using the optimal solver [13], the problem of linear allocation of keywords can be solved.Damie et al.[14] proposed a similar data attack that utilizes a small number of known queries as prior knowledge. Xu et al. [16] proposed an attack in a dynamic scenario, which utilizes the update volume and query response volume of the same keyword before and after the update, and combines query equality to recover the query. Gui et al. [17] redefined leakage from the perspective of SSE systems and pointed out that the vast majority of literature only considers the leakage of encrypted search indexes throughout the SSE system and does not take into account the leakage of encrypted documents. Attacks remain effective

against some SSE schemes that suppress leaks. Wang et al. [18] proposed a machine learning enhancement framework for real-time risk assessment in searchable encryption. In other fields of cybersecurity, machine learning is also widely applied to malware detection [19], [20] and threat detection [21]. In [12], [24], the authors indicate that similar data attacks can bypass the defense of pattern randomization [35], [15]. Additionally, active learning techniques [36] can be utilized to enhance entity recognition, which may be related to keyword extraction in such attacks.

Damie et al. [14] indicated that accuracy is related to query volume (the number of documents matched by the query). Oya et al. [12] pointed out that high-frequency queries are more likely to be accurately recovered. In [7], the authors emphasize that high-volume queries are easier to recover. Nie et al. [22] pointed out that high-volume and high-frequency queries are easier to recover than other queries. Based on these queries, almost all remaining queries can be recovered. The volume and frequency of keywords in the database follow Zipf's law [23], that is, the lower the ranking of a keyword, the smaller its volume and frequency. The similar data attack RSA proposed by Damie et al. [14] can recover all queries by using only 10 known queries. Passive query recovery attacks are classified into two types based on the prior knowledge utilized by the attacks: known data attacks and similar data attacks. Table I lists several known data attacks and similar data attacks.

Inspired by Nie et al.[22], prioritizing the recovery of high-volume and high-frequency queries brings significant advantages to subsequent query recovery. Fig. 1 shows the overview of the refine_atk (Ours) operation pipeline.The main contributions of this study are as follows:

- Most existing query recovery attacks simulate user queries using plaintext queries, whereas this study employs ciphertext queries for user query simulation. The statistical features used throughout the attack are modeled based on ciphertext queries. The volume, frequency, and co-occurrence information of ciphertext queries remain consistent with those in the plaintext state. Since the indices of the underlying semantic embeddings of queries are fixed, semantic vectors can still be extracted via these indices, enabling effective ciphertext query recovery and facilitating better analysis and validation of attack performance.
- This study uses a multi-layer perceptron (MLP) to dynamically allocate weights for volume and frequency. The weights learned by the MLP cannot be directly applied to distinctive query recovery; instead, a coefficient α for distinctive query recovery is derived with the help of a small number of known queries. The weights assigned to volume and frequency, as well as coefficient α vary in each simulated attack.
- This study constructs a TF-IDF weighted co-occurrence matrix from unrecovered queries and unpaired keywords. The pre-trained language model MiniLM-L6 is used to encode unrecovered queries and unpaired keywords to obtain a semantic embedding submatrix. These two components form a cost matrix, and the remaining queries are recovered in one pass via greedy graph matching.
- The proposed method (Ours) remains robust even when the attacker possesses limited auxiliary knowledge. The first module of the proposed attack achieves promising Top-10 performance, meaning that the top 10 candidates in the attack results all correspond to the recovery of distinctive queries, which is critical to the overall effectiveness of our method.

II. SSE Scheme and Attack Model

A. SSE

Searchable encryption schemes are divided into five steps. Generate key: Enter a security parameter and output the system key. The data owner encrypts the file with a key and uploads it to the cloud server. Index establishment: Extract keywords from the database, encrypt as user document indexes, and upload to the cloud server. Generate trapdoor: The data owner uses the key pair to generate trapdoors for the keywords and send to the cloud server. Search: The cloud server matches the search trapdoor with the encrypted index stored locally. If the match is successful, the cloud server will send the ciphertext document mapped by the encrypted index to the user. File decryption: Legitimate users decrypt ciphertext documents.

Bost et al. [25] proposed a forward-privacy SSE scheme named Sophos, which only relies on trapdoor permutation and does not require a structure similar to ORAM. Subsequently, Bost [26] proposed a framework for constructing forward privacy using range-constrained pseudo-random functions. Cash and Tessaro[27] discussed the balance issue between "server storage size" and "memory access locality" in SSE. The size of the encrypted index is O(NlogN), and only O(logN) reads are required during search. Chamani et al. [28] proposed three DSSE schemes, addressing the balance between "security and performance" in backward privacy and providing efficient solutions for strong privacy requirements. Kamara and Moataz [29] proposed a non-interactive and efficient SSE scheme that can handle arbitrary disjunctive queries and boolean queries. It features sublinear search efficiency in the worst-case scenario and optimal communication complexity. Patel et al. [30] proposed a scheme that supports dynamic sharing and unsharing of documents among users, with small cross-user leakage and slow growth as queries increase. Sun et al. [31] proposed the cryptographic primitive of symmetric deletable encryption, which features fast search speed and low server storage cost. Xu et al. [32] introduced the "key-updatable pseudo-random function" to construct the ROSE scheme, which simultaneously satisfies robustness, forward privacy, backward privacy and practical performance.

B. Pattern Leakage

Searchable encryption can leak the document identifier (access pattern) returned by the query. By observing and analyzing the access patterns, attackers attempt to infer the private information queried by users. The search pattern refers to which two queries in a query sequence are equal. SSE using deterministic encryption algorithms will expose the same access pattern when repeatedly searching for the same keyword. Taking advantage of this, attackers can infer whether two queries correspond to the same keyword. The volume pattern refers to the number of documents returned by each query.

TABLE I. COMPARISON OF THE EXISTING PASSIVE ATTACKS[1]

Attack	Leak	KPK		SPK		ACC
		Doc	Query	Doc	Fre	
IKK[6]	AP	■	▲	□	□	∼80%
Count[5]	AP,VP	■	□	□	□	∼90%
SubgraphID[7]	AP	▲	□	□	□	∼90%
LEAP[8]	AP	▲	□	□	□	∼100%
RSA[14]	AP	□	▲	■	□	∼85%
IHOP[24]	AP,SP	□	□	■	■	∼90%
Jigsaw[22]	AP,SP,VP	□	□	■	▲	∼90%
Ours(refine_atk)	AP,SP,VP	□	▲	■	▲	**∼ 95%**

[1] "AP" stands for access pattern, "SP" for search pattern, "VP" for volume pattern, "KPK" for known prior knowledge, "SPK" for similar prior knowledge. "Doc" stands for document, "Query" for known query, and "Fre" for frequency. "■" indicates that the attacker requires all known data or highly depends on similar data. "▲" indicates that the attacker requires some known data or relatively depends on similar data. "□" indicates that the attacker does not need known data or similar data. The first four are known data attacks, and the last four are similar data attacks. Although our attack requires some known queries to assist in attacks, in actual scenarios, it is "treated as" not knowing any known queries.

Searchable encryption can reveal the number of documents returned by the server.

C. Attacker

Honest, yet curious servers follow the SSE protocol, but attempt to recover user queries by exploiting leaks and prior knowledge. The server can access all encrypted documents. In searchable symmetric encryption, only the document contents and indexes are encrypted, while the semantic information of keywords (e.g., word meaning, contextual relations) naturally exists in plaintext space. Although an attacker cannot directly obtain the plaintext of user queries, they can extract the keyword space W_s from auxiliary data documents and independently perform semantic embedding on these keywords. Since semantic embedding relies solely on public corpora and pre-trained models and does not require any user private data, the attacker is fully capable of incorporating this information into the attack process. Relying solely on volume, frequency, and co-occurrence matrices, the attacker struggles to distinguish between queries with similar statistical features (e.g., two low-frequency, low-volume keywords may have nearly identical volume and frequency values). This fully demonstrates the ambiguity inherent in statistical information, whereas semantic embedding provides orthogonal discriminative capability. Query recovery essentially aligns observed ciphertext queries (with observable volume, frequency, and co-occurrence patterns) with plaintext keywords. Semantically similar words often exhibit similar document occurrence patterns (e.g., "apple" and "fruit" may appear in similar document contexts). Therefore, semantic embedding and co-occurrence-based statistical features are naturally complementary.

Attackers use leaks to infer the frequency, volume and co-occurrence information of queries. Suppose a user issues s queries, represented as Q^s, and l different queries are identified by using the search frequency. Mark the query list identified by the attacker as $Q_r = [q_1, q_2,..., q_l]$. The attacker observed the documents returned by the query $[D(q_1), D(q_2),..., D(q_l)]$. The server knows the total number of documents and records it as $|D|$. For each query q, normalize the query volume: $v_{qi} = |D(q_i)| / |D|$, $V_r = [v_{q1}, v_{q2},..., v_{ql}]$ is the volume vector for all queries in the Q_r, with the component numerator being the number of documents matched by the query. The attacker uses the search pattern to know the frequency of q in Q^s. The frequency of q is: $f_q = Cnt(q)/|Q^s|$, where $Cnt(q)$ is the number of q in Q^s. $F_r = [f_{q1}, f_{q2},..., f_{ql}]$ is the frequency vector for all queries in Q_r. Based on the access pattern, an attacker can construct an $l \times |D|$ matrix ID_r, where the rows are queries and the columns are documents. If the search results of q_i include d_j, then $ID_r[i, j] = 1$; otherwise 0. Finally, construct the $l \times l$ co-occurrence matrix $C_r = ID_r ID_r^T / |D|$.

Suppose the attacker has a list of similar documents $D_s = [d_1, d_2,..., d_k]$, and uses the same algorithm to extract the keywords. Attackers can easily obtain a keyword space similar to the user's document[3], denoted as $W_s = [w_1, w_2,..., w_m]$, and construct a volume vector $V_s = [v_{w1}, v_{w2},..., v_{wm}]$, with $v_{wi} = |D_s(w_i)| / |D_s|$, where $|D_s(w_i)|$ is the number of documents containing the keyword w_i in D_s. The attacker constructs an ID_s from W_s and D_s, with the co-occurrence matrix $C_s = ID_s ID_s^T / |D_s|$. For W_s, the attacker obtained the similar search frequency $F_s = [f_{w1}, f_{w2},..., f_{wm}]$ through Google Trends [37], which is used to display the keyword trends and changes searched by users in various regions around the world on the Google search engine.

III. REFINE_ATK

A. Partition Query

Fewer keywords in the database have high volume and frequency. Users may frequently query these keywords, so it is necessary to first distinguish between high-volume and high-frequency queries. Fig. 2 and Fig. 3 shows the results of the queries ranked by volume and frequency. This study selects the top 1000 keywords without stopping words as W. The query is divided into four quadrants: HVHF, HVLF, LVHF and LVLF. "L" and "H" represent "low" and "high", while "V" and "F" represent "volume" and "frequency". High-volume (frequency) queries refer to those ranking in the top 10% in terms of volume (frequency), while low-volume (frequency) queries account for the remaining 90%.

Rank the queries based on volume and frequency, and display the queries in different quadrants according to the ranking. The HVHF query ranking in the upper right corner is high, but the quantity is small. The ranking of the LVLF query in the lower left corner is low, but the number is large.

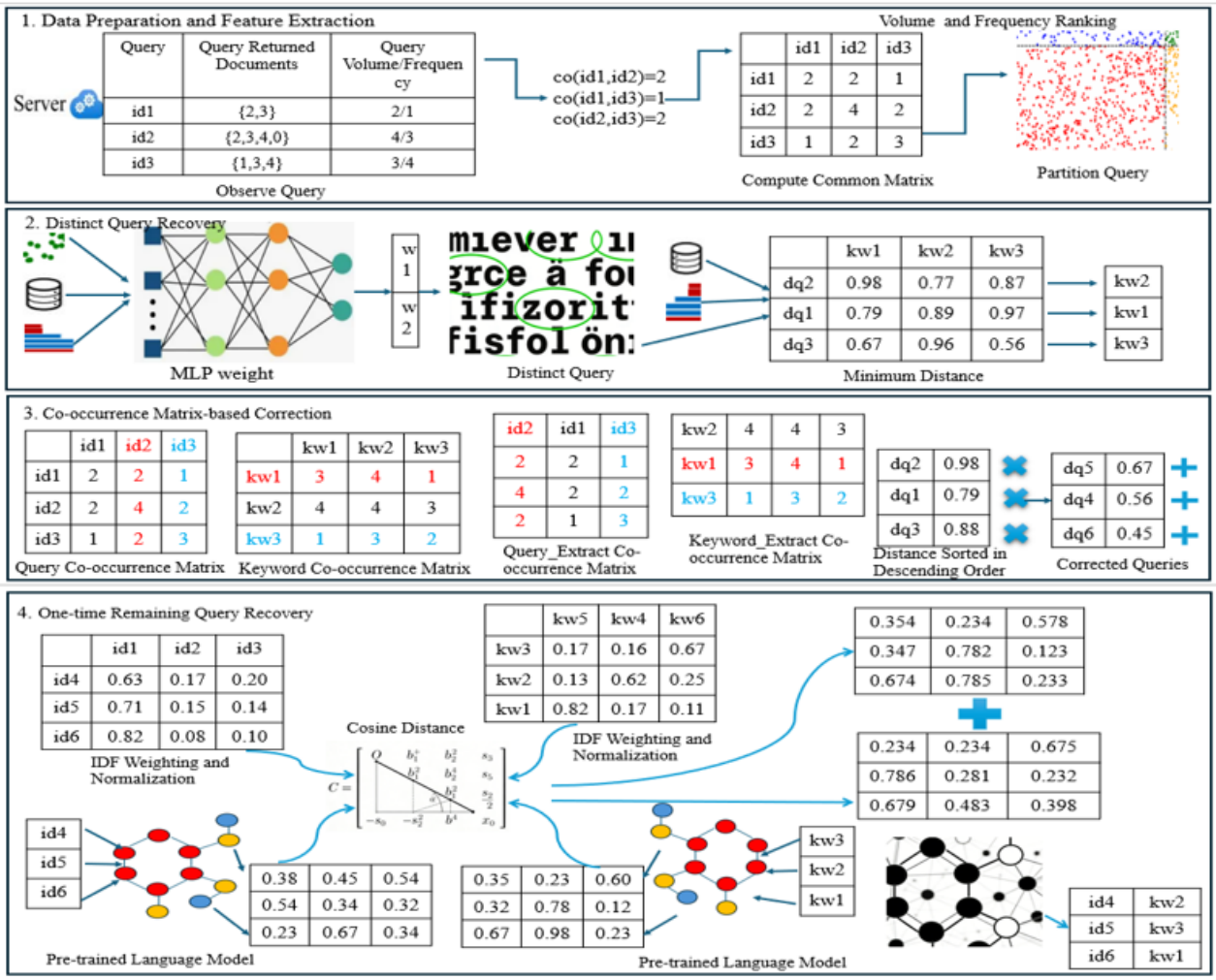


Fig. 1. An overview of the refine_atk (Ours) operation pipeline.

The attack accuracy rates in the HVLF quadrant in Enron and Lucene were 63% and 86%, respectively, 84% and 72% for LVHF, 26% and 34% for LVLF, and all were over 90% for HVHF. It indicates that the LVLF query is difficult to recover.

B. Identify and Recover Distinctive Queries

Algorithm 1 inputs the Td_r, V_r, F_r of the leaked exported data in the SSE query and similar data W_s, V_s, V_s. Output the *BaseRec* predictions. *BaseRec* refers to the number of queries recovered by this module. The *BaseRec* should not be too large. If the *BaseRec* is too large, there will be more predictions for non-HVHF queries, and the accuracy rate will decline.

First, the HVHF, HVLF, LVHF, and LVLF queries are divided. MLP uses the HVHF query, F_r and V_r learn the weights of volume and frequency, which are respectively denoted as w_V and w_F. Transpose F_r and V_r into column vectors and perform a logarithmic transformation (The frequency is processed in the same manner) [see Eq. (1)]:

$$\log V = \ln(V + \epsilon) \tag{1}$$

Element-by-element multiplication and division [see Eq. (2)]:

$$V \times F = V \odot F \tag{2}$$

$$V_{divF} = \frac{V + \epsilon}{F + \epsilon} \tag{3}$$

Calculate the ascending ranking for F_r and V_r, respectively. The smallest element ranks 0, the second smallest ranks 1, and so on. The largest ranking is n-1, where n is the total number of queries. Normalize the ranking vector to the interval [0,1]:

$$\text{ranks}_V = \frac{\text{argsort}(\text{argsort}(V))}{n-1} \tag{4}$$

Concatenate the features V, F, $logV$, $logF$, $V \times F$, V_{divF}, F_{divV} [Swap the numerator and denominator, see Eq. (3)], $ranks_V$ and $ranks_F$ [Replace the numerator with the frequency see Eq. (4)] by column. If the query $\in$ HVHF, the label is marked as 1; otherwise 0. If the number of HVHF queries

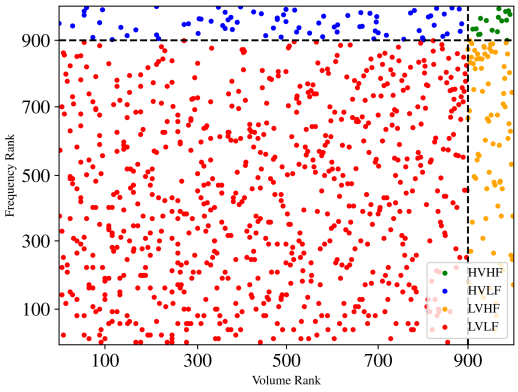


Fig. 2. Query distribution divided into different quadrants based on volume and frequency rank in Enron.

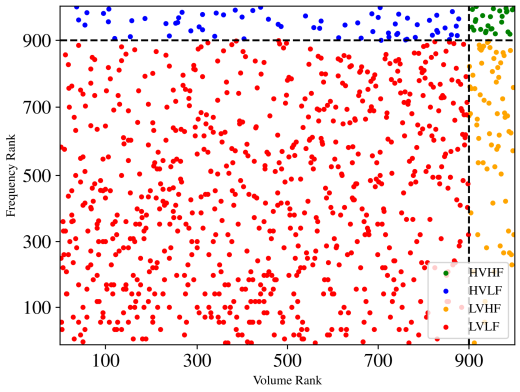


Fig. 3. Query distribution divided into different quadrants based on volume and frequency rank in Lucene.

Algorithm 1 Locate and recover distinctive queries

procedure LocRecDQ($\mathrm{Td}_r, \mathrm{V}_r, \mathrm{F}_r, \mathrm{W}_s, \mathrm{V}_s, \mathrm{F}_s$, params)
// Outline: learn weights using MLP, compute the differential distance, select the distinctive queries, and then perform recovery.Params include: BaseRec, nb_queries.
// 1. Partition the quadrants and take HFHV (or another subset) as the training set.
$(\mathrm{HFHV}, \mathrm{HFLV}, \mathrm{LFHV}, \mathrm{LFLV}) \leftarrow$ plot_histogram_and_classify($\mathrm{F}_r, \mathrm{V}_r$)
// 2. Use MLP to learn the weight vector on the selected subset (output denoted w).
TrainSet $\leftarrow$ HFHV
w $\leftarrow$ learn_weight_mlp(TrainSet, $\mathrm{F}_r, \mathrm{V}_r$)
// 3. Compute the differential distance vector with the learned weights and select the distinctive queries.
$\mathrm{D_{FV}} \leftarrow$ calculate_dVF(w)
// 4. Obtain known query pairs and learn the recovery coefficient α.
known_pairs $\leftarrow$ get_known_queries(nb_queries = params.nb_queries)
$\alpha \leftarrow$ learn_recover_weight_alpha(known_pairs)
// 5. Perform recovery mapping on the distinctive queries.
Pred $\leftarrow$ RecoverByVF(HFHV, $\mathrm{W}_s, \mathrm{V}_s, \mathrm{F}_s, \alpha$)
return Pred
end procedure

is less than 5% of the total number of queries, oversampling is performed. Specifically, the query index for HVHF is pos_idx, and the query index for the rest is neg_idx. The number of HVHF queries is the maximum value between 50 and 10% of the total number of queries. The number of repetitions, repeat_times, is the number of HVHF queries divided by the length of pos_idx, repeat the HVHF query features and labels repeat_times, and concatenate them with the remaining queries. Finally, shuffle the order. Divide the training set and the validation set (80% training and 20% validation). Use StandardScaler (The validation set is treated in the same way) [see Eq. (5)]:

$$X_{\mathrm{train}_r} = \frac{X_{\mathrm{train}} - \mu_{\mathrm{train}}}{\sigma_{\mathrm{train}}} \tag{5}$$

The hidden layer size of MLPClassifier is (64, 32), the activation function is relu, the solver is adam, an early stop strategy is set, and the maximum iteration is 500 times. Repeatedly shuffle a certain feature column on the validation set to see how much the model performance drops. The more the performance drops, the more important this feature is. Each feature is shuffled 20 times and the average is taken. Each feature has an importance score. Aggregate feature importance to *V* and *F*. For each feature *i*, its importance is I_i, and it is assigned according to the feature name (Apply the same processing to the frequency) [see Eq. (6)]:

$$\mathrm{imp}_V = \sum_{\text{only } V} I_i + \sum_{V \text{ and } F} \frac{I_i}{2} + \sum_{\text{others}} \frac{I_i}{2} \tag{6}$$

Only *V* refers to a feature *i*, where the feature name contains only *V*, with an importance of I_i. *V* and *F* indicates that the feature name includes both *V* and *F*, with an importance of $\frac{I_i}{2}$. Others refers to all other features, with an importance of $\frac{I_i}{2}$. The weight calculation for volume and frequency is simplified as follows [see Eq. (7) to Eq. (9)]:

$$w = \left[\frac{\mathrm{imp}_V}{|\mathrm{imp}_V| + |\mathrm{imp}_F|}, \frac{\mathrm{imp}_F}{|\mathrm{imp}_V| + |\mathrm{imp}_F|} \right] \tag{7}$$

$$w_V = \frac{\mathrm{imp}_V}{|\mathrm{imp}_V| + |\mathrm{imp}_F|} \tag{8}$$

$$w_F = \frac{\mathrm{imp}_F}{|\mathrm{imp}_V| + |\mathrm{imp}_F|} \tag{9}$$

Calculate the differential distance to identify unique queries from all the queries. The differential distance d_{qi} of query q_i is [see Eq. (10)]:

$$d_{q_i} = \min_{q_j \in Q_r \wedge j \neq i} w_V \cdot |v_{q_i} - v_{q_j}| + w_F \cdot |f_{q_i} - f_{q_j}| \tag{10}$$

w_V and w_F denote the weights associated with volume and frequency, respectively. The queries are sorted in descending order based on their differential distance. The top *BaseRec* queries constitute the distinctive queries. A small set of known queries is utilized to learn the weight parameter α, which is used to recover distinctive queries. The specific value of α is determined by the statistical characteristics (volume and frequency) of the known queries and their correspondence with keywords in the auxiliary data. Known queries are used solely as a tool for internal parameter estimation and are not directly leveraged as prior knowledge for computing the final attack accuracy, thereby ensuring fairness in comparison with other baseline attacks (e.g., RSA, Jigsaw). This study experimentally validates (see Fig. 20) that with a small number of known queries (1, 2, 5), Algorithm 1 can still achieve relatively stable accuracy (although accuracy fluctuates more significantly when the number is extremely small). This demonstrates that the assumption is reasonable and feasible in practice. Attacker only needs a very small number of random known queries to effectively estimate α.

Following the definition of distance between queries and keywords from[22], each query is paired with its nearest keyword. The distance $s(q_i, w_j)$ between a true query q_i and a keyword $w_j \in W_s$ is defined as Eq. (11):

$$s(q_i, w_j) = \alpha \cdot |v_{q_i} - v_{w_j}| + (1 - \alpha)|f_{q_i} - f_{w_j}| \tag{11}$$

recovery the query q_i as Eq. (12):

$$\text{Pred}(q_i) = \arg \min_{w_j \in W_s} s(q_i, w_j) \tag{12}$$

By dynamically allocating the weights of volume and frequency, this module is capable of recovering unique queries.

C. Correction Through Co-occurrence Matrix

The query-keyword pairs output by the co-occurrence matrix correction of Algorithm 1 is given. The accurate recovery of HVHF queries is crucial for the recovery of subsequent queries.

The input of Algorithm 2 is the co-occurrence matrices C'_r and C'_s, corresponding respectively to the query and keyword in *Pred*, *BaseRec*, and *Pred* recovered by Algorithm 1. That is, in *Pred*, the columns of the query index in the co-occurrence matrix C_r are taken as C'_r, and the rows of the keyword index in the co-occurrence matrix C_s are taken as C'_s. Then, each row of C'_r and C'_s is divided by the sum of the rows to normalize. *ConfRec*($\leq$ *BaseRec*) refers to the number of queries correctly recovered in *Pred*. The smaller the *ConfRec* is, the higher the accuracy rate will be and the fewer queries can be recovered. For the correct prediction (q_i, w_i) in *Pred*, C'_r*[i]* and C'_s*[i]* are similar, and their volume and frequency are similar; Conversely, q_i and w_i are only similar in volume and frequency. The reverse confidence revconf of (q_i, w_i) is Eq. (13):

$$\text{revconf} = \|C'_r[i] - C'_s[i]\| \tag{13}$$

Algorithm 2 Verification based on co-occurrence matrix

procedure VERIFY($Pred, C'_r, C'_s, BaseRec, ConfRec$) //
// Extract the columns and rows corresponding to the queries and keywords in the co-occurrence matrices C_r and C_s of Algorithm 1 $Pred$, respectively, to construct C'_r and C'_s.
  $TempPred \leftarrow Pred$
  $Td' \leftarrow Pred.td$
  $Revconf \leftarrow \emptyset$
  for $i \in [0, |Td'| - 1]$ **do**
    revconf $\leftarrow \|C'_r[i] - C'_s[i]\|$
    append($Td'[i]$, revconf) to $Revconf$
  end for
  sort $Revconf$ in descending order by the revconf value;
  for $i \in [0, BaseRec - ConfRec - 1]$ **do**
    $(td, revconf) \leftarrow Revconf[i]$
    remove the prediction for td from $TempPred$
  end for
  return $TempPred$
end procedure

The greater the *revconf*, less reliable it is for q_i to predict w_i, that is, the plaintext keywords at the bottom of q_i are very likely not w_i. Calculate the *revconf* of each query in *Pred* and add it to the *Revconf*. Sort the *Revconf* in descending order. Remove the first *BaseRec-ConfRec* queries from *Pred*. Return *ConfRec* query predictions.

D. Recovering the Remaining Queries Globally

The input of Algorithm 3 include: the query-keyword pair *Pred* recovered by Algorithm 2, the co-occurrence matrix C_r between the unrecovered and recovered query, all queries T_r, the similar keyword space W_s, and the co-occurrence matrix C_s between the unpaired and paired keyword. U_{td} stands for unrecovered queries, and U_{kw} stands for unpaired keywords.

Calculate TF-IDF weight to each row of $\widetilde{C}_r$ and $\widetilde{C}_s$ [see Eq. (14) and Eq. (15)]:

$$idf(j) = \log \frac{N}{df(j) + 1} \tag{14}$$

$$r_w[j] = r[j] \times idf(j) \tag{15}$$

N represents the number of unrecovered queries, *df* is the number of non-zero elements in each column of the $\widetilde{C}_r$ matrix, *j* is the index, and *r* is the $\widetilde{C}_r$ matrix. Calculate the cosine distance between the two weighted co-occurrence matrices [see Eq. (16)]:

$$D_{co}[i, j] = 1 - \frac{\tilde{C}_r[U_{td}[i]] \cdot \tilde{C}_s[U_{kw}[j]]^T}{\|\tilde{C}_r[U_{td}[i]]\| \cdot \|\tilde{C}_s[U_{kw}[j]]\|} \tag{16}$$

The MiniLM-L6 pre-trained language model has 6 hidden layers and 12 attention heads. The dimension of the hidden layer is 384 dimensions, the maximum sequence length is 128 tokens, and mean pooling is adopted. MiniLM is a

lightweight variant of BERT, with fewer parameters and a more streamlined structure than the standard BERT. Encode the unrecovered queries and unpaired keywords using MiniLM-L6, normalize the generated embedding vectors, and convert them into numpy matrices. In this matrix, the embedding vectors of unrecovered queries and unpaired keywords are concatenated column-wise, respectively, and their indices remain fixed throughout the entire attack process, which is critical to query recovery. For the indexes of unrecovered queries and unpaired keywords, extract the corresponding embedding matrix row vectors to form new embedding submatrices, and calculate the cosine distance between the two embedding submatrices [see Eq. (17)]:

$$D_{emb}[i,j] = 1 - \frac{E_{td}[i] \cdot E_{kw}[j]^T}{\|E_{td}[i]\| \cdot \|E_{kw}[j]^T\|} \tag{17}$$

E_{td} and E_{kw} are embedding submatrices for unrecovered queries and unpaired keywords. The cost matrix is a linear combination of D_{co} and D_{emb} [see Eq. (18)]:

$$M[i,j] = \beta \cdot D_{co}[i,j] + (1-\beta) \cdot D_{emb}[i,j] \tag{18}$$

M is the cost matrix, β is the weight of $D_{co}[i, j]$, and 1-β is the weight of $D_{emb}[i, j]$. This study uses a greedy graph to match unrecovered queries and unpaired keywords. M_{td} and M_{kw} are two empty sets. The row index and column index of matric *M* element, as well as the element values, are added to flat as tuples to sort each element of *M* in ascending order. For each tuple's row index and column index, if neither is in M_{td} or M_{kw}, take the value of the row index in U_{td}, denoted as *td*, and take the value of the column index in U_{kw}, denoted as *kw*. (*td*, *kw*) is added as a matching pair in P_{final} [see Eq. (19)]:

$$P_{final}[td] = kw \tag{19}$$

The M_{td} and M_{kw} collections, respectively add row index and column index [see Eq. (20)]:

$$\begin{cases} M_{td} \leftarrow M_{td} \cup \{i\}, & i \notin M_{td},\ j \notin M_{kw} \\ M_{kw} \leftarrow M_{kw} \cup \{j\}, & j \notin M_{kw},\ i \notin M_{td} \end{cases} \tag{20}$$

The known queries in Algorithm 1 are not the high-volume known queries in RSA [14], but are randomly extracted from the user query sequence and the keyword space.

E. Experimental Data and Evaluation Metrics

The experimental dataset consists of Enron and Lucene. The Enron email Corpus [33] was collected from 2000 to 2002, containing a total of 30,109 emails. The Lucene mailing list consists of 66,491 emails from 2001 to 2020 [34]. This study uses the preprocessed version provided in reference [12].

The search frequency of keywords in Enron and Lucene is generated using Google Trends, which contains Google search trends for approximately 260 weeks from October 2016 to October 2021. The attacker auxiliary knowledge is the sum of the search frequency for each keyword within 1 to 50 weeks. Divide by the total to normalize the frequency of each keyword, denoted as F_s. The user query is simulated and generated based on the total frequency *F* from 1 + τ to 50 + τ. τ is the time gap between the attacker auxiliary knowledge and the user query.

Randomly shuffle all the documents and divide them into two disjoint subsets of equal size. One subset serves as the user's encrypted database D_r, and the other subset as similar data D_s. The attacker generates W_s, V_s and C_s based on D_s, and observes all user queries to obtain Q_r, V_r, F_r and C_r. Each

Algorithm 3 Recover all queries using co-occurrence and embedding.

procedure RECOVERALL($Pred, T_r, W_s, C_r, C_s$)
Input: $Pred$: Algorithm 2 predictions (map td $\rightarrow$ kw); T_r: all trap set; W_s: similar keyword universe; C_r: co-occurrence matrix (rows: unrecovered vs recovered queries); C_s: co-occurrence matrix (rows: unpaired vs paired keywords); $\beta \in [0,1]$, use_emb (flag), optional encoder $\mathcal{E}(\cdot)$.
$P_{final} \Leftarrow Pred$
$U_{td} \Leftarrow T_r \backslash \text{dom}(P_{final})$
$U_{kw} \Leftarrow W_s \backslash \text{range}(P_{final})$
$\widetilde{C}_r \Leftarrow C_r$, $\widetilde{C}_s \Leftarrow C_s$
$idf \Leftarrow$ compute IDF vector over columns of $\widetilde{C}_r$ and $\widetilde{C}_s$
while each row r of $\widetilde{C}_r$ **do**
    $r \Leftarrow r \odot \text{idf}$
    $r \Leftarrow r / \|r\|_2$
end while
while each row s of $\widetilde{C}_s$ **do**
    $s \Leftarrow s \odot \text{idf}$
    $s \Leftarrow s / \|s\|_2$
end while
$D_{co} \Leftarrow$ pairwise cosine-distance between rows of $\widetilde{C}_r[U_{td}]$ and $\widetilde{C}_s[U_{kw}]$
if $use_emb = true$ and $\mathcal{E}$ available **then**
    $E_{td} \Leftarrow \mathcal{E}(U_{td})$
    $E_{td} \Leftarrow$ row-normalize(E_{td})
    $E_{kw} \Leftarrow \mathcal{E}(U_{kw})$
    $E_{kw} \Leftarrow$ row-normalize(E_{kw})
    $D_{emb} \Leftarrow$ pairwise cosine-distance between E_{td} and E_{kw}
else
    $D_{emb} \Leftarrow 0$ (same shape as D_{co})
end if
$M \Leftarrow \beta D_{co} + (1-\beta) D_{emb}$
$\mathcal{M}_{td} \Leftarrow \emptyset$, $\mathcal{M}_{kw} \Leftarrow \emptyset$
$L \Leftarrow$ list of tuples (i, j, M_{ij}) for all valid i, j
sort L ascending by third element (cost)
while each (i, j, ...) in L **do**
    if $i \notin \mathcal{M}_{td}$ and $j \notin \mathcal{M}_{kw}$ **then**
        $id \Leftarrow$ the i-th element of U_{td}
        $kw \Leftarrow$ the j-th element of U_{kw}
        $P_{final}[id] \Leftarrow kw$
        $\mathcal{M}_{td} \Leftarrow \mathcal{M}_{td} \cup \{i\}$
        $\mathcal{M}_{kw} \Leftarrow \mathcal{M}_{kw} \cup \{j\}$
    end if
end while
return P_{final}
end procedure

combination of experimental parameters was independently simulated 30 times. Each simulation randomly selects half of the documents as D_s and generates queries based on the frequency F.

Aggressiveness can be evaluated by accuracy rate, recovery rate and attack time. The recovery rate is the proportion of queries that have been recovered among all queries. Accuracy refers to the queries that have been correctly recovered among the recovered queries. The attack time is the time spent executing Algorithm 1 to Algorithm 3, excluding the time for generating the co-occurrence matrix, query generation, generating the keyword-document matrix, and loading the pre-trained language model [see Eq. (21) to Eq. (23)]:

$$\mathrm{ACC} = \frac{|\mathrm{Recovered}(Q^s)|}{|Q^s|} \tag{21}$$

$$\mathrm{QRR} = \frac{|\mathrm{CorrectRec}(Q^s)|}{|\mathrm{Recovered}(Q^s)|} \tag{22}$$

$$\mathrm{Atk_time} = \mathrm{Alg1.time} + \mathrm{Alg2.time} + \mathrm{Alg3.time} \tag{23}$$

F. Performance of Algorithm 1 and Algorithm 2

In Table II, the parameter experiments are conducted, showing the accuracy, recovery rate and the number of correct recovery of Algorithm 1 and Algorithm 2 in Enron and Lucene when *BaseRec* $\in \{25, 50, 100, 200, 400\}$ and *ConfRec*/*BaseRec* $\in \{1, 0.5, 0.2\}$. part = 1 indicates that Algorithm 2 retains all recovery results of Algorithm 1. The first two columns of data under *ConfRec*/*BaseRec* $\in \{1, 0.5, 0.2\}$ are represented as decimals.

The accuracy rate of various parameter combinations exceeds 65%. As *BaseRec* and *ConfRec*/*BaseRec* decrease, the number of queries recovered by Algorithm 2 decreases, the recovery rate drops, but the accuracy rate increases. In Lucene, when *BaseRec* = 50 and *ConfRec*/*BaseRec* = 0.2, the accuracy rate drops by 1%. When *BaseRec* = 25 and *ConfRec*/*BaseRec* = 0.2, Algorithm 2 achieved recovery rates of approximately 0.66% and 2.04% in Enron and Lucene, respectively, but with extremely high accuracy. The attack proposed in this study outperforms the accuracy and the number of correct recovery achieved in reference [22]. *BaseRec* = 400, *ConfRec*/*BaseRec* = 1, the accuracy rate increased by 23% and 17%, respectively, the number of correct recovery increased by 63 and 57.

Algorithm 1 mainly involves MLP learning the weight of volume and frequency through HVHF query, calculating the differential distance, and identifying unique queries. A coefficient α for recovering unique queries is obtained by using a small number of known queries, and finally the unique queries are recovered by volume and frequency. The attack parameter *BaseRec* represents the total number of queries $|Q_r|$. Each dataset independently simulates 30 experiments, collecting the weights, α , accuracy of Algorithm 1, and attack time of each experiment. The collected weight and α array are divided into buckets with a width of 0.05. The mean and standard deviation of the accuracy within each bucket are calculated. The bucking is done to ensure that all values in both arrays are covered and to average the attack processing time. Test the influence of different weights and α on the accuracy and attack time of Algorithm 1.

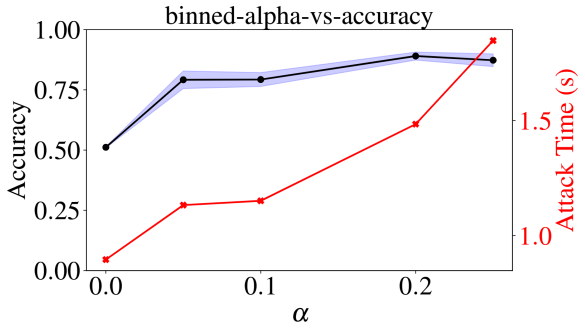


Fig. 4. Accuracy and attack time of Algorithm 1 with different weights and α in Enron.

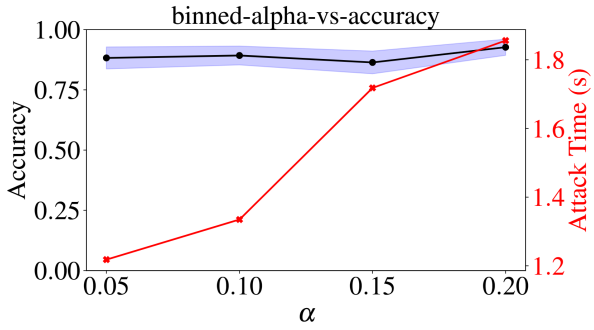


Fig. 5. Accuracy and attack time of Algorithm 1 with different weights and α in Lucene.

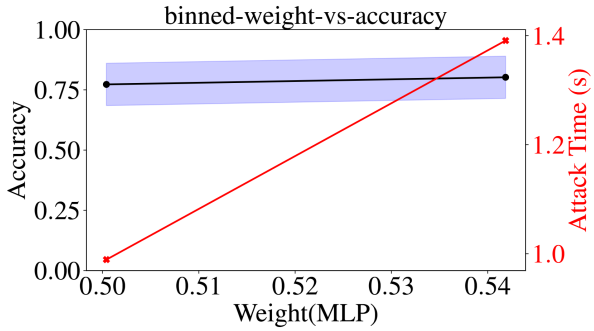


Fig. 6. Accuracy and attack time of Algorithm 1 with different weights and accuracy in Enron.

Fig. 4 shows the results in Enron, where α is 0.2, the accuracy rate is the highest, approximately 87%, and the attack time is about 1.5 seconds. α was selected as 0.05. Fig. 5 shows the results in Lucene, α is selected as 0.05, the attack time is approximately 1.3 seconds, and the accuracy rate is about 87%. In Fig. 6, when the weight keeps changing, the accuracy rate remains basically unchanged. The weight is selected as 0.5, with an accuracy rate of approximately 77% and an attack time of about 1.0 seconds. In Fig. 7, the weight is 0.55, with the highest accuracy rate of approximately 90%, and the

TABLE II. RESULT OF ALGORITHM 1 AND 2 IN ENRON AND LUCENE.

Dataset	BaseRec	part=1			part=0.5			part=0.2		
		ACC	QRR	CRECNUM	ACC	QRR	CRECNUM	ACC	QRR	CRECNUM
Enron	25	0.9942	0.2276	24.4	1.0000	0.0271	12.0	1.0000	0.0066	5.0
	50	0.9600	0.3149	43.3	0.9974	0.0721	24.8	1.0000	0.0178	10.0
	100	0.8986	0.4259	76.5	0.9645	0.1481	45.7	1.0000	0.0453	20.0
	200	0.7779	0.5964	121.6	0.9233	0.2767	82.4	0.9835	0.0840	38.5
	400	0.6728	0.7933	175.2	0.8686	0.4990	130.3	0.9680	0.1602	73.2
Lucene	25	0.9990	0.2835	24.9	1.0000	0.0405	12.0	1.0000	0.0204	5.0
	50	0.9888	0.3547	48.2	0.9997	0.0786	25.0	0.9859	0.0339	9.6
	100	0.9332	0.4747	86.1	0.9788	0.1912	48.3	1.0000	0.0499	20.0
	200	0.8522	0.6417	143.8	0.9897	0.3401	95.5	1.0000	0.1036	40.0
	400	0.7305	0.8301	204.8	0.9292	0.5282	150.5	0.9956	0.2093	78.6

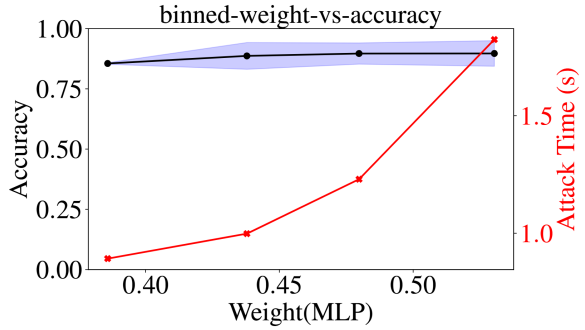


Fig. 7. Accuracy and attack time of Algorithm 1 with different weights and accuracy in Lucene.

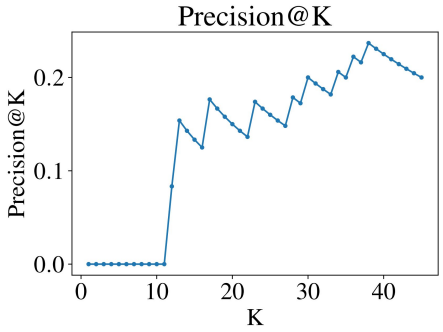


Fig. 8. Result of P_K for Algorithm 1 using grid search for weights in Enron.

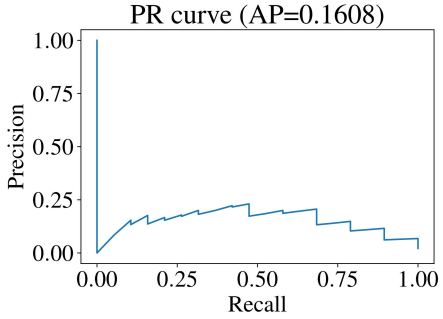


Fig. 9. Result of AP for Algorithm 1 using grid search for weights in Enron

attack time is about 1.8 seconds. The MLP can model the relationship between volume and frequency statistical features, thereby better adapting to the distribution discrepancy between the auxiliary dataset and the target dataset. The learned weights can dynamically adjust the contribution proportions of volume and frequency features, thus improving the generalization performance of the attack under different query distributions. In practical attack scenarios where the attacker only possesses incomplete auxiliary knowledge — for example, when the scale of the auxiliary document set is only half that of the target set — the MLP can mine latent correlations from limited data and maintain high query recovery accuracy.

Algorithm 1 conducts a comparative experiment of grid search and MLP learning volume and frequency weight to identify and recover unique queries. *BaseRec* is 45. Collect the differential distance. Manually divided HVHF queries are the actual queries to be recovered, as well as the prediction results of Algorithm 1 and F_r, V_r. Test the influence of grid search weight and MLP learning weight on Algorithm 1.

Fig. 8 and Fig. 11 are the Precision@K result graphs of the grid search weights of Algorithm 1 in Enron and Lucene. Algorithm 1, which only utilizes the grid search weight to recover the results of unique queries in Top-K, is indicated to be not ideal. Algorithm 1 focuses on how many of the top K of its prediction results are correct recoveries for unique queries. Because Algorithm 1 has a poor ability to recover unique queries, it will affect the performance of the entire attack algorithm. Algorithms with high classification accuracy but low Top-K accuracy are of no use to attacks. Fig. 9 and Fig. 12 show the PR curves in Enron and Lucene respectively. The average precision AP in Enron is 0.1608, and that in Lucene is 0.4058. Fig. 10 and Fig. 13 show the ROC curves in Enron and Lucene. The AUC value in Enron is 0.9204, and that in Lucene is 0.9712. In [22], the authors indicate that the keyword universe size is 1000, and the number of unique queries is approximately 18. Positive samples (unique queries) are extremely unbalanced, while PR and ROC curves are more sensitive and meaningful to the problem of sample imbalance.

Fig. 14 and Fig. 17 are the Precision@K result graphs

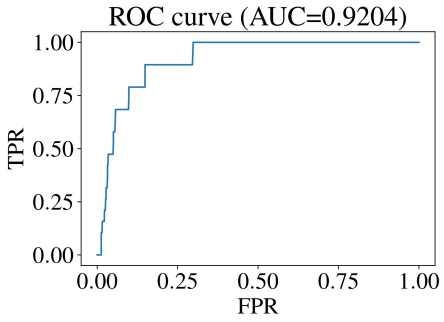


Fig. 10. Result of AUC for Algorithm 1 using grid search for weights in Enron

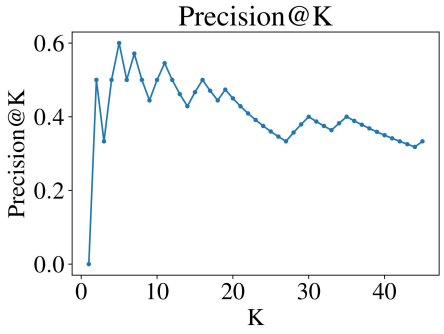


Fig. 11. Result of P_K for Algorithm 1 using grid search for weights in Lucene

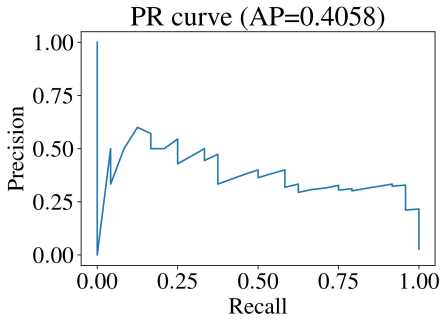


Fig. 12. Result of AP for Algorithm 1 using grid search for weights in Lucene

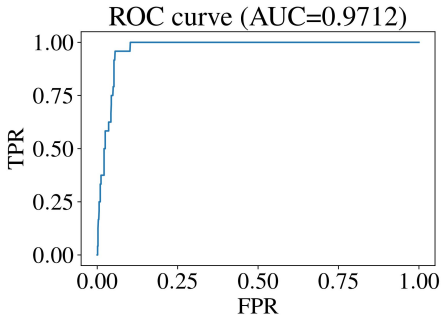


Fig. 13. Result of AUC for Algorithm 1 using grid search for weights in Lucene

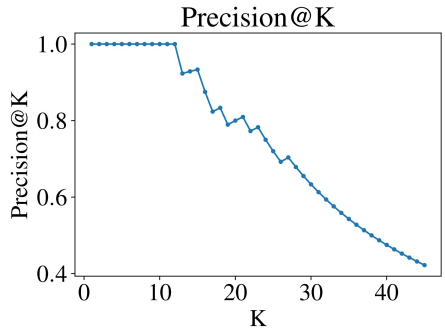


Fig. 14. Result of P_K AUC for Algorithm 1 using MLP to learn weight in Enron

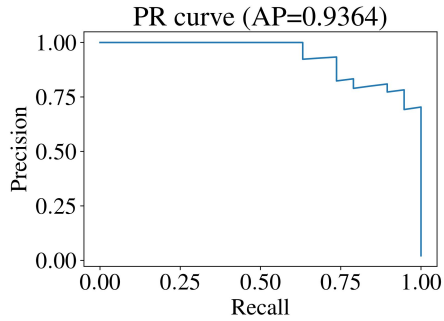


Fig. 15. Result of AP for Algorithm 1 using MLP to learn weight in Enron

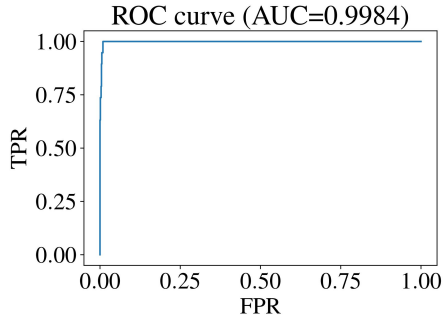


Fig. 16. Result of AUC for Algorithm 1 using MLP to learn weight in Enron

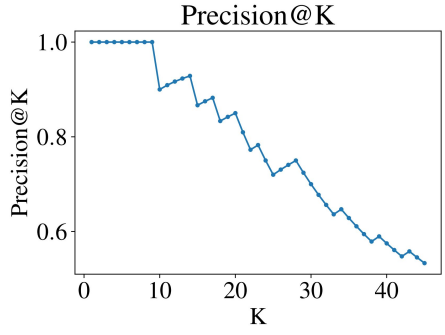


Fig. 17. Result of P_K for Algorithm 1 using MLP to learn weight in Lucene

of Algorithm 1 using MLP to learn the weights of volume and frequency in Enron and Lucene. In Enron and Lucene,

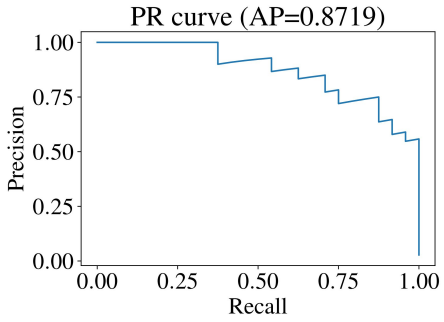


Fig. 18. Result of AP for Algorithm 1 using MLP to learn weight in Lucene

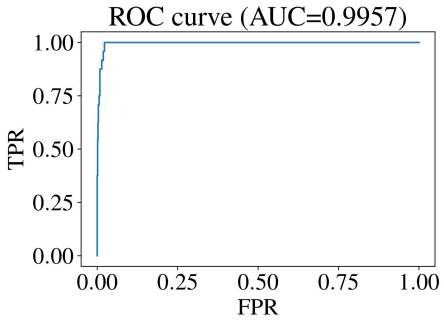


Fig. 19. Result of AUC for Algorithm 1 using MLP to learn weight in Lucene

the accuracy of K at 10 is nearly 1, Under Lucene, when K is between 0 and 10, the highest accuracy of the grid search method is only 0.6 (Fig. 11), while the accuracy of the MLP method is nearly 1 (Fig. 17). By comparing, it can be concluded that, regardless of which dataset, when K is any value within the interval, the MLP method is much more accurate than the grid search method. The accuracy of the MLP method in the Top-10 is nearly 1, significantly revealing that the Top 10 recovery results of Algorithm 1 are all recoveries of unique queries. The AP values in Enron and Lucene are 0.9364 (Fig. 15) and 0.8719 (Fig. 18), respectively. The AUC values are 0.9957 (Fig. 19) and 0.9984 (Fig. 16). In terms of AP values, the MLP method outperformed the grid search method by 83% and 53%, respectively, AUC values were 8% and 2% higher.

Algorithm 1 uses a small number of known queries to obtain a coefficient α for recovering the target query, and finally recovers the target query using volume and frequency. Test the impact of the known number of queries on the accuracy and attack time of Algorithm 1. Collect the accuracy rate and attack time of Algorithm 1 for each experiment under different known query numbers.

Fig. 20, when it is known that the number of queries is 1 and 2, the accuracy of Algorithm 1 is not stable. The lower quartile differs significantly from the upper quartile, and the data is highly dispersed and fluctuates greatly. Excluding outliers, the highest accuracy rate is 92%. In Fig. 21, when the number of known queries is small, the main reason for the unstable accuracy is that the selection of the Algorithm 1 coefficient α depends on the known queries. When the number

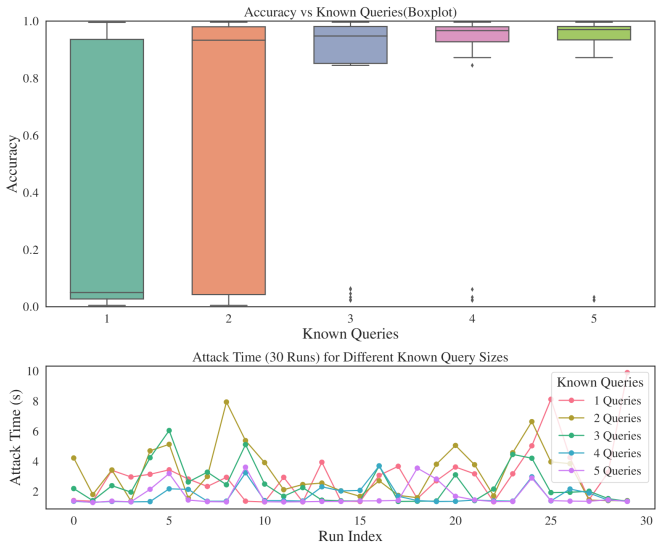


Fig. 20. Accuracy and per-attack time of Algorithm 1 with different known queries in Enron.

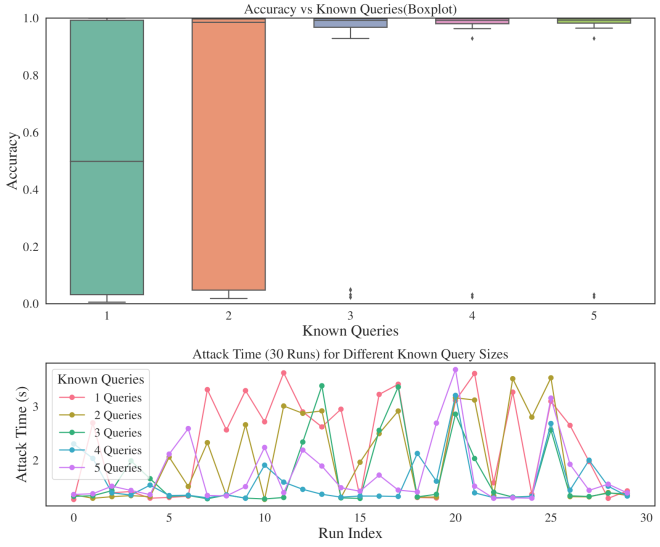


Fig. 21. Accuracy and per-attack time of Algorithm 1 with different known queries in Lucene.

is small, it is insufficient to support the selection of coefficients, resulting in unstable accuracy of Algorithm 1. The specific coefficients are shown in Fig. 4 and Fig. 5.

G. Attack Results and Durability

The following are the attack results of the entire algorithm and the impact of β on accuracy. *BaseRec* is 100 and *ConfRec* is 50. The size of the keyword space is 1000, the number of queries observed each week is 2000, and the total observation period is 50 weeks.

When $\beta = 0$, query recovery only relies on the embedding sub-matrix. Although it has a relatively high accuracy rate, in order to utilize the knowledge of weighted co-occurrence

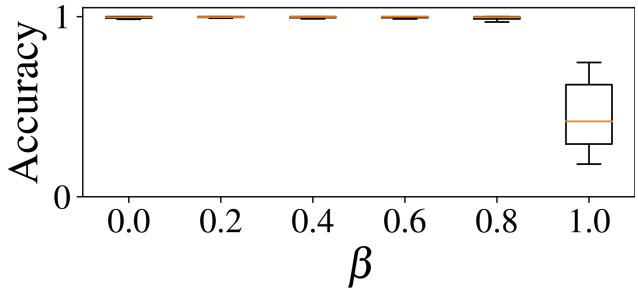


Fig. 22. Accuracy of the refine_atk under different β with embedding matrices in Enron.

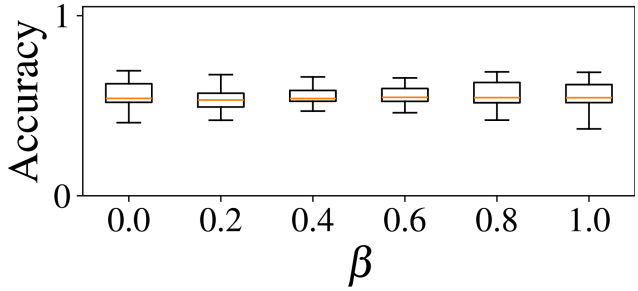


Fig. 23. Accuracy of the refine_atk under different β without embedding matrices in Enron.

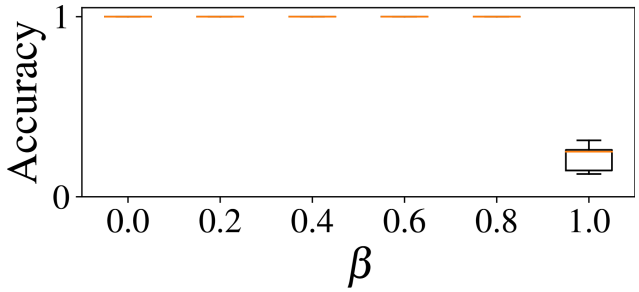


Fig. 24. Accuracy of the refine_atk under different β with embedding matrices in Lucene.

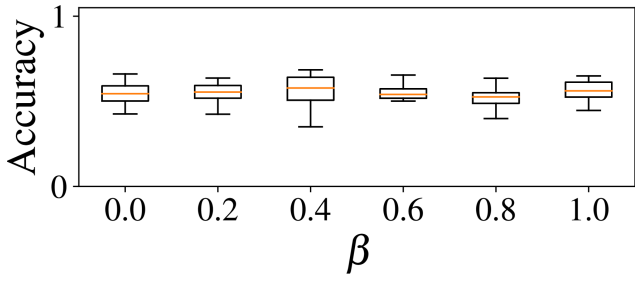


Fig. 25. Accuracy of the refine_atk under different β without embedding matrices in Lucene.

matrices, β is not set to 0. This indirectly reflects the significance of embedding sub-matrices for query recovery. Fig. 22 and Fig. 24 show the experimental results of the β value in Algorithm 3. When β = 1, query recovery only relies on the weighted co-occurrence matrix, with accuracy rates of 40% and 30% in Enron and Lucene, respectively. As β is reduced, by introducing an embedding sub-matrix, the accuracy rate is improved. The accuracy rate in Enron and Lucene is as high as 98%. This indicates that the embedding sub-matrix is crucial to Algorithm 3 and significantly enhances the attack performance. Without the embedding sub-matrix, the attack accuracy rate drops to around 50% [Fig. 23 and Fig. 25].

The embedding sub-matrix was introduced in Algorithm 3. The absence of an embedding sub-matrix will not affect the performance of Algorithm 1 and Algorithm 2, and unique queries can still be identified and recovered with high accuracy.

TABLE III. RESULT OF ATTACK ACCURACY FOR OUTDATED FREQUENCY UNDER DIFFERENT τ VALUES (IN WEEKS).

Dataset	τ (in weeks)			
	$\tau = 10$w	$\tau = 50$w	$\tau = 100$w	$\tau = 150$w
Enron	0.9879	0.9581	0.9272	0.9273
Lucene	0.9995	0.9774	0.9991	0.9890

The impact of "outdated" frequency on attack accuracy. "Outdated" refers to the time difference between the auxiliary frequency and the actual frequency. If the frequency information of the two overlaps less, the attack may fail. Durability refers to whether an attack can recover future queries by leveraging "outdated" frequency. Table III presents the experimental results for measuring the durability of attacks. The experiment used the frequency of Google Trends [37] from 1 to 50 weeks as the auxiliary information for attackers, and the frequency of 1 + τ to 50 + τ weeks was used to simulate the generation of user queries.

When τ is 10 and 50, the attack accuracy rates in Enron and Lucene decrease by 3% and 2%, respectively. When τ is 50 and 100, Enron drops by 3% and Lucene rises by 2%. When τ is 100 and 150, the attack accuracy rate in Enron remains basically unchanged, while it decreases by 1% in Lucene. The results show that the attack proposed in this study can still successfully recover future queries by leveraging the auxiliary frequency from a long time ago.

H. User Query Distribution

The impact of different query distribution on attack accuracy. The distribution for simulating the generation of user queries is "real", "poisson", "uniform", and "zipfian". "real" refers to the simulation of user query generation using the frequency of 1 + τ to 50 + τ weeks in Google Trends. "poisson" refers to generating query records that follow a Poisson distribution, simulating the random characteristic of "the number of times a keyword is queried within a unit of time". "uniform" refers to a uniform probability distribution, meaning that the probability of all keywords being queried is exactly the same, simulating the ideal situation where "users pay the same attention to all keywords". "zipfian" refers to generating queries that follow the Zipf distribution, which conforms to the practical rule that "a few keywords are frequently queried while the majority are infrequently queried". The attack parameters *BaseRec* is 45, *ConfRec* is 35, and β is 0.5. The size of the experimental parameter keyword space is 1000, the number of queries observed each week is 500, the observation period is 50, τ is 50, and the number of known queries is 15. The distribution of user queries generated

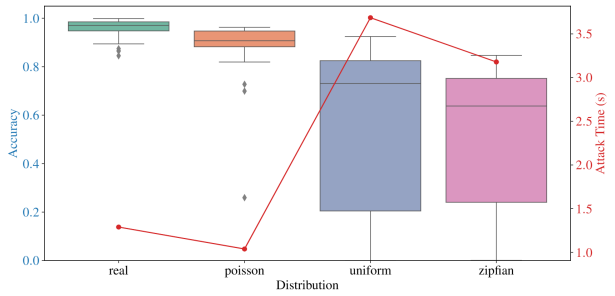


Fig. 26. Accuracy and average attack time of refine_atk under different distributions in Enron.

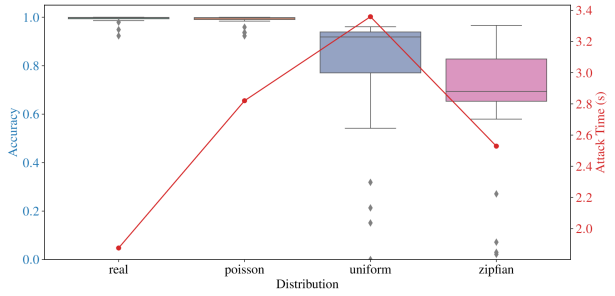


Fig. 27. Accuracy and average attack time of refine_atk under different distributions in Lucene.

is "real", "poisson", "uniform", and "zipfian". The following shows the results of the entire algorithm in Enron and Lucene.

In Fig. 26, the accuracy rate of the "real" distribution is as high as 97%. The average attack time of the "poisson" distribution is the shortest, at 1.1 seconds. The accuracy of "uniform" and "zipfian" is unstable because the queries generated by these two distributions contain a large number of difficult-to-recover LVLF queries. The average attack time of the "uniform" distribution is the longest, at 3.7 seconds. Fig. 27, the accuracy rates of the "real" and "poisson" distributions are 99%. The accuracy rates of "uniform" and "zipfian" are unstable, but perform better than Enron, with the lowest accuracy rate exceeding 60%. This is because the contents of the Lucene and Enron datasets are not related. Additionally, Lucene has a relatively large proportion of HVHF queries. The average attack time of the "uniform" distribution is the longest. The distribution that generates user queries has a certain impact on accuracy.

Table IV presents the results of paired t-tests, which compare the attack accuracy of the proposed refine_atk and the baseline Jigsaw Attack across two datasets (Enron and Lucene) under two keyword distributions (uniform and Zipfian). The core metrics include mean value, standard deviation, mean difference, p-value, Cohen's d effect size, and 95% confidence interval (CI).

Under the uniform distribution on the Enron dataset, refine_atk achieves an average accuracy of 0.9587, which is significantly higher than 0.8521 of Jigsaw Attack, with a mean improvement of 0.1067. Statistical tests confirm that the difference is statistically significant (p = 0.0012 $<$ 0.05), with a Cohen's d of 0.6548, indicating a medium-to-large effect size. The 95% CI [0.0458, 0.1675] does not contain zero, which further verifies the significance of the difference. The standard deviation of refine_atk is only 0.0074, much lower than 0.1620 of Jigsaw Attack, demonstrating that the attack results of refine_atk are extremely stable with negligible fluctuation across repeated experiments in this scenario.

The refine_atk achieves an average accuracy of 0.9678, substantially outperforming Jigsaw Attack (0.7261) with a mean improvement of 0.2417 under the uniform distribution on Lucene. The difference is highly statistically significant (p $<$ 0.0001), with a Cohen's d of 1.0842, which corresponds to a large effect size and indicates a pronounced performance gap between the two attacks. The 95% CI [0.1584,0.3249] does not contain zero, validating the statistical significance. Meanwhile, the standard deviation of refine_atk is only 0.0112, while that of Jigsaw Attack reaches 0.2261, highlighting the remarkable robustness advantage of the proposed method.

Jigsaw Attack achieves an average accuracy of 0.7894, which is significantly higher than 0.6370 of refine atk under the Zipfian distribution on Enron, with a mean difference of -0.1524. The difference is highly statistically significant (p $<$ 0.0001), with a Cohen's d of -0.8897, representing a large effect size. The 95% CI [-0.2163, -0.0884] does not contain zero, confirming the significance of the difference. The standard deviation of refine atk is 0.1706, much higher than 0.0593 of Jigsaw Attack, indicating that the attack results of refine atk fluctuate greatly in this scenario and its stability is inferior to the baseline method.

The accuracy of the two methods is nearly comparable under Zipfian distribution on Lucene: with a mean difference of only -0.0049. Statistical tests show that $p = 0.8438$, which is far greater than 0.05, indicating no statistically significant difference between the two methods. The Cohen's d is -0.0363, representing a negligible effect size, which means the performance of the two attacks is essentially equivalent in this scenario. With the exception of the Lucene-Zipfian combination, the remaining groups of comparisons reach statistical significance with effect sizes of medium level or above. This result demonstrates that the observed performance differences are not caused by random errors, but are genuine gaps derived from the inherent characteristics of the algorithms.

IV. Attack Experiment and Performance Analysis

This study will compare the performance of the refine_atk (Ours) attack with the Jigsaw attack [22], RSA attack [14], and IHOP attack [24].

A. Experimental Settings and Attack Parameters

The keyword space is the top $|W|$ keywords in terms of volume ranking. Evaluate the above attacks for different $|W|$ values (500, 1000, 2000). Suppose the attacker knows the frequency of each keyword in the keyword space in the first 50 weeks. By leveraging the frequency of 50 to 100 weeks in Google Trends, generate η queries per week, with τ = 50. η is 100, 500 and 2500.

Briefly introduce the attack parameters required for the above-mentioned attack and the reasons for their values. The

TABLE IV. STATISTICAL RESULTS OF ATTACK ACCURACY COMPARISON[1]

Dataset	Distribution	Refine_atk		Jigsaw Attack		Mean Diff.	p-value	Cohen's d	95% CI	
		Mean	Std	Mean	Std				Lower	Upper
Lucene	zipfian	0.6904	0.1242	0.6953	0.0534	-0.0049	0.8438	-0.0363	-0.0553	0.0455
Lucene	uniform	0.9678	0.0112	0.7261	0.2261	0.2417	< 0.0001	1.0842	0.1584	0.3249
Enron	zipfian	0.6370	0.1706	0.7894	0.0593	-0.1524	< 0.0001	-0.8897	-0.2163	-0.0884
Enron	uniform	0.9587	0.0074	0.8521	0.1620	0.1067	0.0012	0.6548	0.0458	0.1675

[1] All experiments are conducted with a keyword universe size of 1000 and 30 iterations.

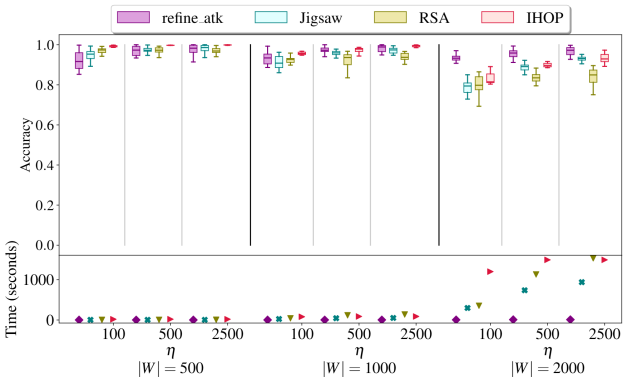


Fig. 28. Comparison of accuracy and time under different attacks in Enron.

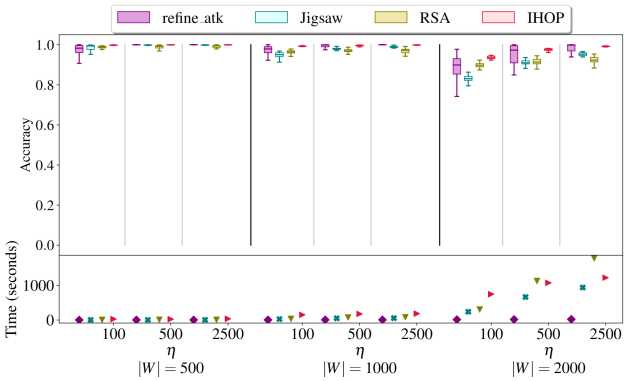


Fig. 29. Comparison of accuracy and time under different attacks in Lucene.

selection of β, *BaseRec* and *ConfRec* significantly affects the accuracy of refine_atk attacks. For β, select a value between 0.2 and 0.8 (Fig. 22). Fig. 6 has shown that without the embedding sub-matrix, the attack accuracy rate drops to 50%. Consider embedding sub-matrices and choose a relatively β, such as 0.5. The minimum *ConfRec* of Jigsaw [22] is 5. If *ConfRec* < 5, Algorithm 2 only correctly recovers < 5 queries, so that Algorithm 3 almost has to recover all user queries. Algorithm 1 and Algorithm 2 do not offer any favorable conditions for the query recovery task, which goes against the original design intention of the entire algorithm. *BaseRec* and *ConfRec* are 45 and 35, respectively, consistent with Jigsaw [22]. The refine_atk is known to have 15 queries. However, during the experiment, these queries are still regarded as "unknown".

For other attack parameters, for Jigsaw attack [22], α is 0.3, β is 0.9. *RefSpeed* controls the number of queries recovered in each iteration. If there are defense measures, increase the *RefSpeed* to reduce the attack time and improve the accuracy. RSA is known to have 10 queries, that is, it randomly selects 10 queries and reveals the underlying plaintext keywords. $|W| \leq 2000$, *RefSpeed* is 10; $|W| > 2000$, *RefSpeed* is 50 [22]. The *RefSpeed* parameter setting strategy for RSA is consistent with that of Jigsaw. The p_{free} of IHOP is 0.25 and that of n_{iters} is 500.

In Fig. 28, the attack accuracy of refine_atk(Ours) is comparable to that of Jigsaw and IHOP. when $|W|$ increases, the attack accuracy rates of Jigsaw, RSA and IHOP decline, but the refine_atk can still remain $> 90\%$. This is because a larger $|W|$ will introduce more low-volume queries that are difficult to recover. The attack time of refine_atk does not fluctuate too much with the changes of $|W|$ and η, and always remains stable. $|W| = 2000$, the attack time of Jigsaw is greater than 500 seconds, and RSA and IHOP are greater than 1000 seconds. This is disadvantageous in some cases because when the attack lasts too long, some databases or systems will record such attack behaviors and issue alerts.

In Fig. 29, when $|W| = 2000$ and $\eta = 100$, the attack accuracy of IHOP is even higher than the refine_atk. The fundamental reason is the extreme sparsity of statistical information. In this case, IHOP's global iterative Hungarian algorithm recovers the global structure from sparse co-occurrence more effectively than refine_atk stagewise greedy strategy, and refine_atk suffers from a severe shortage of HVHF seeds. However, in practical application scenarios, refine_atk outperforms IHOP and other existing attacks by leveraging semantic embeddings, efficient stagewise recovery, and strong robustness. The experimental results in the study confirm this (in Fig. 29, under the vast majority of parameter combinations, refine_atk achieves an accuracy of $\geq 90\%$, and its attack time is far lower than that of IHOP).

B. Refine_atk Ablation Study

1) Known queries ablation experiment: Algorithm 1 recovers unique queries with the aid of known queries. This study now investigates the impact of known queries on the attack performance of Algorithm 1, as shown in Fig. 30. We conducted 30 independent attack simulations on the Enron and Lucene datasets. The keyword space size is $|W|$=1000, *BaseRec*=1000, *ConfRec*= None, the number of observed queries per week $\eta = 5000$, the observation period is 30 weeks, $\tau = 0$, and the user query distribution is "real". In Fig. 30, "enron-kq" denotes the experiment on the Enron dataset with the aid of known queries, while "enron" denotes the experiment without known queries.

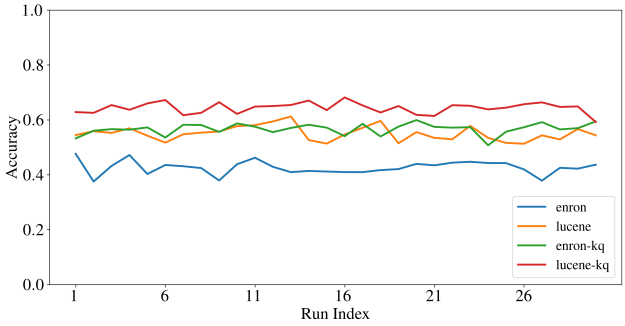


Fig. 30. Ablation study of the known queries component in Algorithm 1.

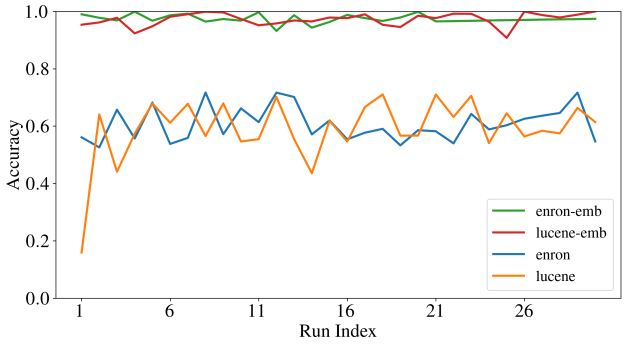


Fig. 31. Ablation study of the semantic embedding component in Algorithm 3.

The accuracy of enron-kq improved by approximately 0.1–0.15 over enron, increasing from 0.4–0.5 to 0.5–0.6, while lucene-kq improved by about 0.1 over lucene, from 0.5–0.6 to 0.6–0.7. It can be observed that the accuracy achieved with known queries is not very high. This is because when *BaseRec*=1000, Algorithm 1 recovers 1000 queries as unique queries, whereas in practice, for a keyword space size of 1000, there are only about 18 unique queries [22].*BaseRec*=1000 far exceeds the actual number of unique queries, leading to the low accuracy of Algorithm 1. Therefore, a lower *BaseRec* value can be set to improve accuracy; please refer to Section III-F for the value of *BaseRec*.

2) Semantic ablation experiment: As shown in Fig. 31, Algorithm 3 in Section III-D recovers the remaining queries at one time with the help of semantic embedding. This section investigates the impact of semantic embedding on the overall attack algorithm. We conducted 30 independent attack experiments on the Enron and Lucene datasets, respectively. The experimental settings are as follows: the keyword space size $|W| = 1000$, $\beta = 0.5$, *BaseRec*=100, *ConfRec*=25, the number of observed queries per week $\eta = 5000$, the number of observed weeks is 30, $\tau = 0$, the number of known queries is 15, and the user query distribution is "real". In Fig. 31, "enron-emb" denotes the experiment with semantic embedding on the Enron dataset, and "enron" denotes the experiment without semantic embedding.

After introducing semantic embedding, the accuracy of "enron-emb" and "lucene-emb" is almost stable between 0.95 and 1.0, while the attack accuracy without semantic embedding is only between 0.4 and 0.7, which is much lower than that

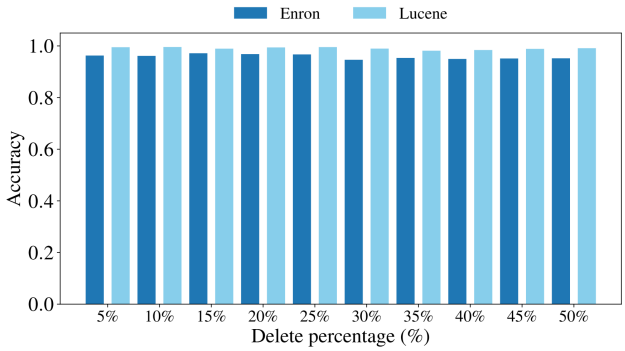


Fig. 32. Comparison of accuracy and delete percentage in Enron and Lucene.

of the attack with semantic embedding. Semantic embedding is a key factor to improve attack accuracy, as it provides additional semantic information for query recovery. The attack performances of the Enron and Lucene datasets show almost no difference, indicating that the attack method with semantic embedding has good generalization to different datasets. The initial fluctuation of Lucene is obvious, which is related to the keyword distribution of the dataset.

C. Delete Similar Data

An experiment on the impact of deleting some similar data on attack accuracy and the recovery of the top 10 high-frequency queries under the partial deletion ratio of Algorithm 1. Deleting partially similar data refers to the assumption that the size of similar data documents is 50 and the deletion ratio is 50%, that is, 50% of the similar data documents are retained, which is 25 documents. The attacker only used these 25 documents to generate the keyword-document matrix, while the real data documents always maintained the size after division. Under the partial deletion ratio, record the number of occurrences of each query in all pellations in the recovery results of Algorithm 1. Each query is sorted in descending order of its occurrence frequency as a list of high-frequency queries. Filter out the pairings that "the query is not on the high-frequency list" or "the predicted keyword is not on the high-frequency list", and retain the top 10 valid pairings where both parties belong to the high-frequency list. The deletion ratio is 5%, ..., 50%. Thirty experiments were independently simulated under each dataset and deletion ratio, and the accuracy of each experiment was collected. Each round of the simulated attack is conducted independently, and there is no mutual interference between different test runs.

When the deletion ratio is 5%, 10%, 15%, 20% and 25%, the attack accuracy rate is slightly higher than that of other ratios. This rule is reflected in both datasets [Fig. 32]. The accuracy rate of the deletion ratio of 5%, 10%, 15%, 20% and 25% in Enron was 96%. The accuracy rate of deletion ratios of 30%, 35%, 40%, 45% and 50% in Lucene is 97%. Regardless of the deletion ratio, the attack accuracy rate is all $>$ 95%. The number of auxiliary documents used by the attacker to recover the query is only half of the user's real documents. Under such unfavorable circumstances, the refine_atk attack can still achieve an accuracy rate of $>$ 90%. In practical scenarios, attackers typically have only limited

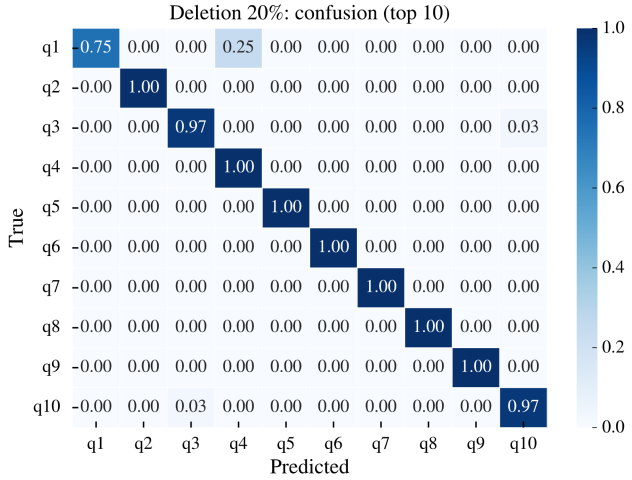


Fig. 33. Recovery results of the top 10 high-frequency queries by Algorithm 1 under 20% ratio. The vertical axis represents the labels of real queries.The horizontal axis represents the labels of the corresponding predicted keywords. (q1, q3) indicates that the real label q3 is predicted as q1.

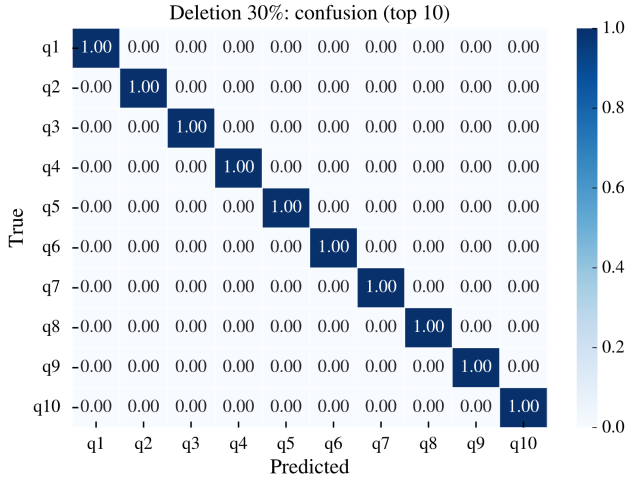


Fig. 34. Recovery results of the top 10 high-frequency queries by Algorithm 1 under ratio 30%.

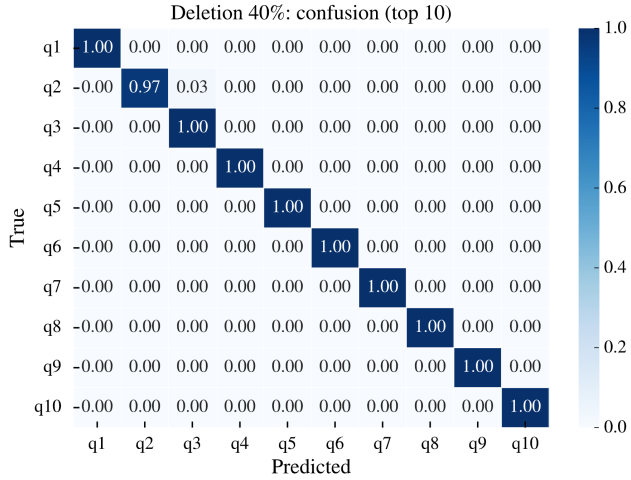


Fig. 35. Recovery results of the top 10 high-frequency queries by Algorithm 1 under ratio 40%.

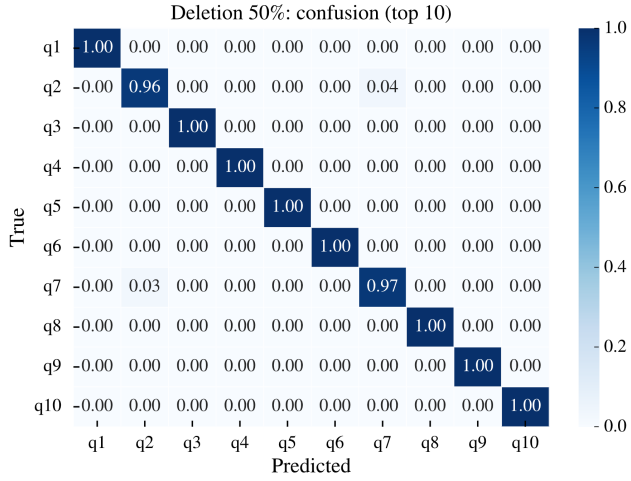


Fig. 36. Recovery results of the top 10 high-frequency queries by Algorithm 1 under ratio 50%.

auxiliary knowledge. Even under this restrictive condition, the proposed attack can still successfully recover users' underlying keywords, which alerts users that their query behaviors entail potential privacy leakage risks.

This shows the recovery of the top 10 high-frequency queries under the partial deletion ratio of Algorithm 1. Fig. 33 shows that the probability of correct recovery for query q1 is relatively low, at 0.75. Recover to q4 with a probability of 0.25. q3 and q10 were correctly recovered with a probability of 0.97. Fig. 34 shows that the probability of correct recovery for query q1-q10 is relatively high, nearly 1. Fig. 35 shows that the probability of correct recovery for query q2 is 0.97. Fig. 36 shows that queries q2 and q7 were correctly recovered with probabilities of 0.96 and 0.97 respectively. This proves the experimental conclusion in Fig. 32, that is, when the attacker auxiliary knowledge is relatively limited, the attack proposed in this study can still correctly recover high-frequency queries with a probability of nearly 1. This suffices to demonstrate the robustness of Algorithm 1 and its sensitivity to HVHF queries. Algorithm 1 exhibits sound generalization performance on both datasets.

V. Conclusion

The proposed attack, refine_atk, performs excellently in most cases. The Algorithm 3 encodes unrecovered queries and unpaired keywords with the help of a pre-trained language model. The refine_atk is tested on different datasets, and the attack accuracy is around 95%. The attack time is much lower than other attacks. Compared with Jigsaw [22], under different datasets and parameters of *BaseRec* and *ConfRec*, the refine_atk achieves higher attack accuracy, more correctly recovered queries, and a recovery rate comparable to that in [22].Algorithm 1, in this study can significantly improve the attack accuracy in Top-10, and can even utilize the weights of query learning capacity and frequency such as LVLF (HVLF, LVHF). The attack performance remains robust when auxiliary

knowledge is limited. In defense evaluation, more advanced defense mechanisms such as semantic obfuscation or synonym expansion should be considered.

The query recovery attack can be extended to dynamic searchable symmetric encryption (DSSE) scenarios. Most existing query recovery attacks are modeled on static document collections, whereas dynamic update operations, such as document addition, deletion and modification, are ubiquitous in practical cloud storage environments and introduce additional novel leakage sources, including update patterns and temporal access correlations. Future work can leverage the leakage characteristics inherent in the dynamic update process to evaluate the query privacy leakage risks of state-of-the-art DSSE schemes in dynamic scenarios. Therefore, it is crucial to develop a secure and practical SSE scheme that can resist semantic-level attacks to protect user data security.

ACKNOWLEDGMENT

This study was supported by the Key Project of the University-Level Scientific Research Innovation Project of Gansu University of Political Science and Law (Grant No. GZF2025XZD27). The authors would like to express their sincere gratitude to the funding organization for its support of this research.

REFERENCES

[1] Y. Zhang, J. Katz, C. Papamanthou (2016) All Your Queries Are Belong to Us: The Power of File-Injection Attacks on Searchable Encryption. Paper presented at the 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, Aug 2016, 707–720. https://doi.org/10.1109/EuroSP48549.2020.00030.

[2] R. Poddar, S. Wang, J. Lu, R. A. Popa (2020) Practical Volume-Based Attacks on Encrypted Databases. Paper presented at the IEEE Symp Security Privacy (S&P), 2020, 354–369. https://doi.org/10.1109/EuroSP48549.2020.00030.

[3] X. Zhang, W. Wang, P. Xu, L. T. Yang, K. Liang (2023) High Recovery with Fewer Injections: Practical Binary Volumetric Injection Attacks against Dynamic Searchable Encryption. Paper presented at the 32nd USENIX Security Symposium (USENIX Security 23), 2023, 5953–5970.

[4] L. Zhang, J. Wang, J. Wu, Y. Wang, S. Sun (2025) Violin: Powerful Volumetric Injection Attack Against Searchable Encryption with Optimal Injection Size. IEEE Trans Industr Inf 21 (9):4103–4115. https://doi.org/10.1109/TDSC.2025.3543248.

[5] D. Cash, P. Grubbs, J. Perry, T. Ristenpart (2015) Leakage-Abuse Attacks Against Searchable Encryption. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2015, 668–679. https://doi.org/10.1145/2810103.2813700.

[6] M. S. Islam, M. Kuzu, M. Kantarcioglu (2012) Access Pattern disclosure on Searchable Encryption: Ramification, Attack and Mitigation. Paper presented at the Netw Distrib Syst Security Symp (NDSS), 2012.

[7] L. Blackstone, S. Kamara, T. Moataz (2020) Revisiting Leakage Abuse Attacks. Paper presented at the Netw Distrib Syst Security Symp (NDSS), 2020.

[8] J. Ning, X. Huang, G. S. Poh, J. Yuan, Y. Li, J. Weng, R. H. Deng (2021) LEAP: Leakage-Abuse Attack on Efficiently Deployable, Efficiently Searchable Encryption with Partially Known Dataset. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2021, 2307–2320. https://doi.org/10.1145/3460120.3484540.

[9] Y. Zhou, Y. Ren, M. Yi, Y. Xiao, Z. S. Tan, N. Moustafa, Z. Tian (2023) CDTier: A Chinese Dataset of Threat Intelligence Entity Relationships. IEEE Trans Sustain Comput 8 (4):627–638. https://doi.org/10.1109/TSUSC.2023.3240411.

[10] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, Z. Tian (2023) CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution. IEEE Trans Knowl Data Eng 35 (6):5695–5709. https://doi.org/10.1109/TKDE.2022.3175719.

[11] C. Liu, L. Zhu, M. Wang, Y. Tan (2014) Search Pattern Leakage in Searchable Encryption: Attacks and New Construction. Inf Sci 265:176–188. https://doi.org/10.1016/j.ins.2013.11.021.

[12] S. Oya, F. Kerschbaum (2021) Hiding the Access Pattern is Not Enough: Exploiting Search Pattern Leakage in Searchable Encryption. Paper presented at the 30th USENIX Security Symposium (USENIX Security 21), 2021, 127–142.

[13] M. L. Fredman, R. E. Tarjan (1987) Fibonacci heaps and their uses in improved network optimization algorithms. J ACM 34 (3):596–615. https://doi.org/10.1145/28869.28874.

[14] M. Damie, F. Hahn, A. Peter (2021) A Highly Accurate Query-Recovery Attack against Searchable Encryption using Non-Indexed Documents. Paper presented at the 30th USENIX Security Symposium (USENIX Security 21), 2021, 143–160.

[15] Z. Shang, S. Oya, A. Peter, F. Kerschbaum (2021) Obfuscated Access and Search Patterns in Searchable Encryption. Paper presented at the 28th Netw Distrib Syst Security Symp (NDSS), 2021. https://doi.org/10.14722/ndss.2021.23041.

[16] L. Xu, L. Zheng, C. Xu, X. Yuan, C. Wang (2023) Leakage-Abuse Attacks Against Forward and Backward Private Searchable Symmetric Encryption. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2023, 3003–3017. https://doi.org/10.1145/3576915.3623085.

[17] Z. Gui, K. G. Paterson, S. Patranabis (2023) Rethinking Searchable Symmetric Encryption. Paper presented at the IEEE Symp Security Privacy (S&P), 2023, 1401–1418. https://doi.org/10.1109/SP46215.2023.10179460.

[18] L. Wang, L. Xu, Y. Chen, Y. Zou, C. Wang (2025) ALERT: machine learning-enhanced risk estimation for databases supporting encrypted queries. Paper presented at the 34th USENIX Security Symposium (USENIX Security 25), 2025.

[19] Y. Chai, J. Qiu, L. Yin, L. Zhang, B. B. Gupta, Z. Tian (2022) From Data and Model Levels: Improve the Performance of Few-Shot Malware Classification. IEEE Trans Netw Service Manage 19 (4):4248–4261. https://doi.org/10.1109/TNSM.2022.3200866.

[20] Y. Chai, L. Du, J. Qiu, L. Yin, Z. Tian (2023) Dynamic Prototype Network Based on Sample Adaptation for Few-Shot Malware Detection. IEEE Trans Knowl Data Eng 35 (5):4754–4766. https://doi.org/10.1109/TKDE.2022.3142820.

[21] Z. Wang, Y. Zhou, H. Liu, J. Qiu, B. Fang, Z. Tian (2024) ThreatInsight: Innovating Early Threat Detection Through Threat-Intelligence-Driven Analysis and Attribution. IEEE Trans Knowl Data Eng 36 (12):9388–9402. https://doi.org/10.1109/TKDE.2024.3474792.

[22] H. Nie, W. Wang, P. Xu, X. Zhang, L. T. Yang, K. Liang (2024) Query Recovery from Easy to Hard: Jigsaw Attack against SSE. Paper presented at the 33rd USENIX Security Symposium (USENIX Security 24), 2024, 2599–2616.

[23] G. K. Zipf (1948) Human Behavior and the Principle of Least Effort.

[24] S. Oya, F. Kerschbaum (2022) IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization. Paper presented at the 31st USENIX Security Symposium (USENIX Security 22), 2022, 2407–2424.

[25] R. Bost (2016) $\sum o\phi o\varsigma$: Forward Secure Searchable Encryption. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2016, 1143–1154. https://doi.org/10.1145/2976749.2978303.

[26] R. Bost, B. Minaud, O. Ohrimenko (2017) Forward and Backward Private Searchable Encryption from Constrained Cryptographic Primitives. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2017, 1465–1482. https://doi.org/10.1145/3133956.3133980.

[27] D. Cash, S. Tessaro (2014) The Locality of Searchable Symmetric Encryption. Paper presented at the Advances Cryptology (EUROCRYPT), 2014, 351–368. https://doi.org/10.1007/978-3-642-55220-5_20.

[28] J. G. Chamani, D. Papadopoulos, C. Papamanthou, R. Jalili (2018) New Constructions for Forward and Backward Private Symmetric Searchable Encryption. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2018, 1038–1055. https://doi.org/10.1145/3243734.3243833.

[29] S. Kamara, T. Moataz (2017) Boolean Searchable Symmetric Encryption with Worst-Case Sub-linear Complexity. Paper presented at the Advances Cryptology (EUROCRYPT), 2017, 94–124. https://doi.org/10.1007/978-3-319-56617-7_4.

[30] S. Patel, G. Persiano, K. Yeo (2018) Symmetric Searchable Encryption with Sharing and Unsharing. Comput Security 74:207–227. https://doi.org/10.1007/978-3-319-98989-1_11.

[31] S. Sun, X. Yuan, J. K. Liu, R. Steinfeld, A. Sakzad, V. Vo, S. Nepal (2018) Practical Backward-Secure Searchable Encryption from Symmetric Puncturable Encryption. Paper presented at the ACM SIGSAC Conf Comput Commun Security (CCS), 2018, 763–780. https://doi.org/10.1145/3243734.3243782.

[32] P. Xu, W. Susilo, W. Wang, T. Chen, Q. Wu, K. Liang, H. Jin (2022) ROSE: Robust Searchable Encryption With Forward and Backward Security. IEEE Trans Inf Forensics Security 17:1115–1130. https://doi.org/10.1109/TIFS.2022.3155977.

[33] W. W. Cohen, MLD (2015) Enron email datasets. https://www.cs.cmu.edu/ ./enron/.

[34] Apache Foundation (1999) Mail archives of lucene. https://mail-archives.apache.org/modmbox/#lucene.

[35] G. Chen, T. Lai, M. K. Reiter, Y. Zhang (2018) Differentially Private Access Patterns for Searchable Symmetric Encryption. Paper presented at the IEEE Conf Comput Commun (INFOCOM), 2018, 810–818. https://doi.org/10.1109/INFOCOM.2018.8486381.

[36] Y. Liu, J. Hu, Z. Chen, X. Wan, T.-H. Chang (2023) EASAL: Entity-Aware Subsequence-Based Active Learning for Named Entity Recognition. Paper presented at the AAAI Conf Artif Intell (AAAI), 2023, 1000–1008. https://doi.org/10.1609/aaai.v37i7.26069.

[37] Google (2004) Google trends.

Multiplicative Gate State Space Models with Skip-Net for High Accuracy COVID-19 Time-Series Prediction

Krung Sinapiromsaran, Supakit Sroynam
Faculty of Science-Department of Mathematics and Computer Science,
Chulalongkorn University, Bangkok, Thailand

***Abstract*—The rapid propagation of the COVID-19 pandemic has placed unprecedented strain on global healthcare systems, creating an urgent need for accurate forecasting to optimize resource allocation and policy implementation. However, the highly non-linear and chaotic behavior of infection rates poses significant challenges for traditional statistical and standard deep learning models. This study proposes the Multiplicative Gating State Space Model with skip connection (MG-SSM-s), a novel architecture designed to capture complex temporal dependencies in epidemiological time series. Drawing inspiration from partial autocorrelation, the model extends modern State Space Models (SSMs) by incorporating a learnable multiplicative side channel to dynamically modulating input processing. We evaluated the efficacy of MG-SSM-s using the Google COVID-19 Open Data repository, analyzing daily confirmed cases across 40 countries, including major epicenters such as the USA, India, and Brazil. Using a 30-day look-back window, the proposed model was benchmarked against four baseline architectures: LSTM, Bi-LSTM, GRU, and standard SSM. Performance verification based on Mean Squared Error (MSE) demonstrates that MG-SSM-s outperforms all deep learning baselines and achieves competitive accuracy with a tuned ARIMA model, demonstrating comparable statistical performance to the latter. These results highlight the framework's robustness and potential as a versatile tool for time-series forecasting.**

Keywords*—State Space Model; COVID-19; time-series forecast; univariate time-series

I. Introduction

The COVID-19 pandemic exposed critical vulnerabilities in public health forecasting infrastructure. As infection rates accelerated, healthcare systems faced severe resource constraints, particularly in managing hospital bed capacity, ventilator supplies, and staffing shortages. A primary barrier to effective management was the volatility of daily case counts, which severely hindered proactive capacity planning. Consequently, developing robust methods to forecast epidemiological time-series has become essential for long-term pandemic preparedness.

COVID-19 transmission dynamics exhibit highly non-linear, non-stationary behaviors that challenge traditional statistical modeling. Infection rates rely on multiplicative interactions between transmission rates, susceptible population densities, and contact frequencies [1]. Moreover, sudden shocks, including policy interventions, viral variants, and behavioral shifts, trigger abrupt regime changes that standard time-series models often misclassify as noise [2]. While machine learning architectures have advanced time-series forecasting, simultaneously capturing non-linear multiplicative dynamics, high-frequency volatility, and long-range temporal dependencies remains an open challenge. To address these limitations, we propose the Multiplicative Gating State Space Model with skip connection (MG-SSM-s), an architecture designed specifically for volatile epidemiological data.

MG-SSM-s extends modern State Space Models with a learnable multiplicative gating mechanism and skip connections, achieving a separation between global trends and volatility. The design is inspired by autocorrelation concepts, where multiplicative product interactions enable the model to capture dependencies across multiple temporal lags. This process is similar to how partial autocorrelation identifies significant historical correlations, while skip connections preserve sudden shocks that linear state transitions would otherwise attenuate. By maintaining parallel pathways for trend and volatility, MG-SSM-s retains both long-range awareness and sensitivity to sudden shifts critical for pandemic forecasting.

A. Contributions

The primary contributions of this research are fourfold:

- Novel Hybrid Architecture: MG-SSM-s extends standard State Space Models by incorporating a learnable multiplicative gating mechanism and skip connection, enabling non-linear expression of temporal dependencies.
- Theoretically-Inspired Design: The multiplicative gate is inspired by autocorrelation concepts, enabling the model to dynamically identify and weight relevant temporal patterns, with an explicit skip connection preserving immediate input shocks.
- Comprehensive Evaluation: We evaluate MG-SSM-s on COVID-19 case forecasting datasets from 40 countries against established baselines, including LSTM, Bi-LSTM, GRU, and vanilla SSMs, demonstrating superior performance on non-stationary epidemic data.
- Practical Applicability for Epidemic Forecasting: The architecture is designed for univariate time series forecasting with linear computational complexity, making it suitable for real-time deployment across diverse healthcare systems during active pandemics.

The remainder of this study is organized as follows: Section II reviews related literature in epidemiological forecasting and State Space Models, highlighting the research gaps motivating MG-SSM-s. Section III defines the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). Section IV details the architectural design of the proposed model, and Section V evaluates its performance on COVID-19 datasets. Finally, Section VI concludes the study and outlines future research directions.

II. Related Work

A. Traditional Statistical Methods in Epidemic Forecasting

Early COVID-19 forecasting efforts relied heavily on statistical linear models, most notably the Autoregressive Integrated Moving Average (ARIMA) and its variants [3]. ARIMA models operate by transforming non-stationary data into a stationary series through differencing, then predicting future values based on a linear combination of past errors and past values. While ARIMA provides mathematical interpretability and works well for simple linear trends, it fundamentally struggles with the chaotic volatility and non-linear dynamics of pandemic data. Its reliance on linear assumptions and stationarity makes it ill-suited for capturing the abrupt infection waves and complex environmental interactions characteristic of COVID-19 propagation.

In a comprehensive analysis, it was demonstrated that meteorological factors, such as temperature, relative humidity, and wind speed, exhibit nonlinear correlations with daily new COVID-19 cases across different regions [4]. This nonlinearity persists even when controlling for confounding factors, indicating that pandemic transmission follows non-linear mechanistic principles that linear statistical models cannot express.

B. Deep Learning Approaches for Epidemic Time-Series Forecasting

To address ARIMA's limitations, Deep Learning (DL) approaches were widely adopted for COVID-19 prediction [5], with the Long Short-Term Memory (LSTM) network developed by [6] emerging as a standard baseline. LSTMs overcome limitations of standard RNNs through a memory cell state modulated by three distinct gates (input, forget, output), allowing selective retention or discarding of information over long sequences [7]. In COVID-19 forecasting, LSTMs have demonstrated the ability to learn non-linear infection trends that ARIMA misses [8]; however, they often struggle to capture very sharp, high-frequency changes (sudden spikes from variants or policy shifts) due to the slow adaptation of recurrent cell states.

1) Bidirectional LSTMs (Bi-LSTM): Bidirectional LSTM architectures, such as those introduced by [9], enhance context awareness by processing data bidirectionally. By concatenating hidden states from both directions, the model gains temporal features surrounding each time step. While bidirectional flow improves interpolation and noise smoothing, it substantially increases computational costs and does not fundamentally solve the modeling of immediate, shock-like volatility in real-time forecasting scenarios where future data is unavailable.

2) Gated Recurrent Units (GRU): The Gated Recurrent Unit, proposed by [10], offers a more computationally efficient alternative by merging the cell state and hidden state and combining input and forget gates into a single update gate. This simplification enables faster training and convergence, particularly advantageous when modeling multiple countries simultaneously. However, the GRU shares a fundamental limitation with LSTM: reliance on dense recurrence that becomes opaque and difficult to optimize for extremely long historical contexts, often leading to performance degradation when the look-back window is extended significantly.

C. State Space Models for Time-Series Forecasting

Recently, State Space Models (SSMs) have gained prominence as a powerful alternative to Transformers and RNNs for time-series tasks. The Structured State Space Sequence (S4) model introduced by [11], [12] pioneered this resurgence by combining continuous-time representations from control systems with deep learning. S4 utilizes a low-rank HiPPo matrix initialization to compress infinite-context history into a finite state, enabling efficient handling of extremely long-range dependencies via the convolution theorem during training. However, while S4 excels at capturing global trends over long horizons, its linear time-invariant (LTI) nature can be too rigid for non-stationary pandemic systems undergoing rapid structural changes, such as virus mutations or abrupt policy shifts.

More recently, an advancement to the SSM paradigm was introduced via a selection mechanism that allows system matrices to be functions of the input [13]. This formulation makes the model time-varying and input-selective, which enables it to distinguish between relevant and irrelevant information at each time step, thereby solving the content-aware limitations of S4.

However, directly deploying an unconstrained vanilla Mamba block onto raw, univariate time-series data introduces architectural vulnerabilities. Because standard selective SSMs are optimized to compress dense, multi-channel representations, applying them to low-dimensional, short-horizon vectors yields severe over-parameterization and rapid latent state saturation.

III. Background Knowledge

A. Autocorrelation and Partial Autocorrelation

Time-series forecasting relies fundamentally on analyzing the dependency of a variable on its past values. Two statistical measures are paramount in characterizing these temporal dependencies: the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF).

1) Autocorrelation Function (ACF): The ACF quantifies the linear relationship between a time-series X_t and a lagged version of itself X_{t-k}, accounting for both direct influence and indirect influence propagated through intermediate time steps $(X_{t-1}, \ldots, X_{t-k+1})$. Formally, for a stationary time-series, the autocorrelation coefficient at lag k, denoted as ρ_k, is defined as:

$$\rho_k = \frac{\text{Cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t)\text{Var}(X_{t-k})}} = \frac{\gamma_k}{\gamma_0} \tag{1}$$

where, γ_k is the autocovariance at lag k and γ_0 is the variance. In the context of deep learning, Recurrent Neural Networks (RNNs) naturally model this accumulated dependency through their hidden state updates.

2) Partial Autocorrelation Function (PACF):: While the ACF formulation in Eq. (1) captures total correlation, it fails to isolate the specific contribution of X_{t-k} independent of intervening lags. The PACF addresses this limitation by measuring the correlation between X_t and X_{t-k} after removing the linear dependence explained by the intermediate lags $\{X_{t-1}, \ldots, X_{t-k+1}\}$. Mathematically, the partial autocorrelation ϕ_{kk} at lag k is expressed as the correlation of the residuals in Eq. (2):

$$\phi_{kk} = \text{Corr}(X_t - \hat{X}_t, X_{t-k} - \hat{X}_{t-k}) \tag{2}$$

where, $\hat{X}_t$ and $\hat{X}_{t-k}$ represent the linear projections of X_t and X_{t-k}, respectively, onto the space spanned by the intermediate set $\{X_{t-1}, \ldots, X_{t-k+1}\}$.

The distinction between the total correlation in Eq. (1) and the isolated dependency in Eq. (2) inspires the proposed MG-SSM-s architecture. Standard State Space Models capture past context through recursive state equations. However, to capture dependencies, where X_{t-k} directly influences X_t without intermediate smoothing, a mechanism analogous to the PACF is required. The proposed skip-connection and multiplicative gating mechanism serve this purpose by providing a direct channel to model specific lagged dependencies that standard recurrence might otherwise oversmooth.

B. State Space Model (SSM)

The State Space Model (SSM) serves as a foundational framework for modeling dynamic systems in control theory and signal processing. In the context of deep learning, it allows for the parameterization of dependencies in sequence data through a latent state representation. A Continuous-Time Latent State Model defines a system, where an input signal $u(t) \in \mathbb{R}^D$ drives a latent state $x(t) \in \mathbb{R}^N$, which is subsequently projected to an observable output $y(t) \in \mathbb{R}^D$ [14].

Mathematically, the dynamics of this system are governed by the linear ordinary differential equations (ODEs) defined in Eq. (3) and Eq. (4):

$$x'(t) = \mathbf{A}x(t) + \mathbf{B}u(t) \tag{3}$$
$$y(t) = \mathbf{C}x(t) + \mathbf{D}u(t) \tag{4}$$

The system is parameterized by the state evolution matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ in Eq. (3), which governs the internal dynamics of the hidden state, and the input projection matrix $\mathbf{B} \in \mathbb{R}^{N\times D}$. The latent state is mapped to the observed data via the output projection matrix $\mathbf{C} \in \mathbb{R}^{D\times N}$ in Eq. (4). The final term, $\mathbf{D} \in \mathbb{R}^{D\times D}$, represents a direct feedthrough connection from input to output. In modern structured state space architectures, such as S4 and Mamba, this feedthrough term is frequently omitted from the core ODE dynamics and instead implemented as a separate, learnable residual connection to facilitate optimization.

1) Discretization and Recurrent Form: To facilitate the processing of discrete data sequences, the continuous-time dynamics must be discretized. Modern SSM architectures typically adopt the Zero-Order Hold (ZOH) principle, which posits that the input signal $u(t)$ remains constant within the sampling interval Δ. The discretization process transforms the continuous parameters $(\mathbf{A}, \mathbf{B})$ into their discrete counterparts $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$. The discrete state transition matrix is defined via Eq. (5):

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}) \tag{5}$$

While the input projection matrix $\bar{\mathbf{B}}$ is derived via integration over the sampling period, as shown in Eq. (6):

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \approx \mathbf{A}^{-1}(\bar{\mathbf{A}} - \mathbf{I})\mathbf{B} \tag{6}$$

Consequently, the system can be expressed as a linear recurrence via Eq. (7) and Eq. (8), highlighting the structural parallel between discretized SSMs and Recurrent Neural Networks (RNNs):

$$x_t = \bar{\mathbf{A}}x_{t-1} + \bar{\mathbf{B}}u_t \tag{7}$$
$$y_t = \mathbf{C}x_t \tag{8}$$

In standard Linear Time-Invariant (LTI) SSMs (e.g., S4), the matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ in Eq. (7) are constant across all time steps. This stationarity allows for efficient parallel processing but limits the model's ability to adapt its dynamics to specific input tokens. Recent advancements in structured state-space models, most notably Mamba [13], address the limitations of static content compression by introducing a selection mechanism. By defining the system matrices $(\mathbf{A}, \mathbf{B}, \Delta)$ as functions of the input, these models effectively transform the architecture from an LTI system into a time-varying linear recurrence. While this input-selectivity significantly enhances the model's ability to distinguish relevant context within long sequences, it fundamentally alters the computational dynamics, as the loss of time-invariance.

IV. Methodology

A. Model Structure

The Multiplicative Gating State Space Model with skip connection (MG-SSM-s) is defined by the system of equations parameterized by the learnable set $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{J})$ in Eq. (9) to Eq. (11):

$$h_t = \mathbf{A}h_{t-1} + \mathbf{B}x_t \tag{9}$$
$$g_t = (\mathbf{E}g_{t-1}) \odot \tilde{x}_t + \mathbf{F}x_t \tag{10}$$
$$y_t = \mathbf{C}h_t + \mathbf{D}x_t + \mathbf{J}g_t \tag{11}$$

where, $\odot$ denotes the element-wise (Hadamard) product. The term $\tilde{x}_t \in \mathbb{R}^m$ represents the scalar input x_t projected into the latent gate dimension via the all-ones vector $\mathbf{1}_m \in \mathbb{R}^m$, such that $\tilde{x}_t = x_t \cdot \mathbf{1}_m$. This operation ensures that the

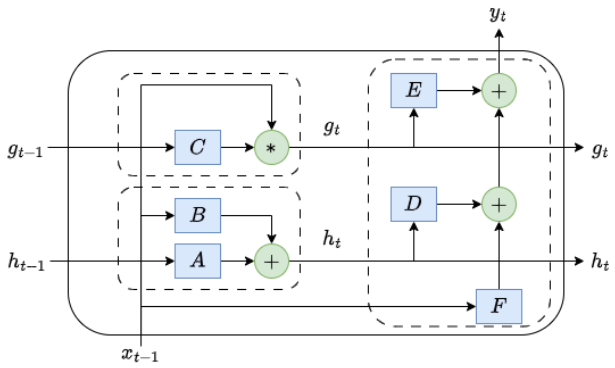


Fig. 1. Architecture of the Multiplicative Gating State Space Model with skip connection (MG-SSM-s).

scalar input modulates each channel of the latent gate state g_t uniformly. Fig. 1 presents the architecture of the Multiplicative Gating State Space Model with skip connection (MG-SSM-s).

The model processes a univariate input sequence $X = (x_1, \ldots, x_T) \in \mathbb{R}^T$ to produce an output sequence $Y = (y_1, \ldots, y_T) \in \mathbb{R}^T$. The internal memory is decoupled into two distinct pathways: a main linear state $h_t \in \mathbb{R}^l$ (where, l is the latent dimension) evolving via Eq. (9), and a multiplicative gate state $g_t \in \mathbb{R}^m$ (where, m is the gate dimension) governed by Eq. (10).

The linear state h_t in Eq. (9) evolves via standard LTI dynamics using the transition matrix $\mathbf{A} \in \mathbb{R}^{l\times l}$ and input projection $\mathbf{B} \in \mathbb{R}^{l\times 1}$. The gated state g_t in Eq. (10) introduces non-linearity through a two-step mechanism: first, the previous state g_{t-1} is transformed by $\mathbf{E} \in \mathbb{R}^{m\times m}$ and modulated element-wise by the input; second, a direct skip connection parameterized by $\mathbf{F} \in \mathbb{R}^{m\times 1}$ is added. This skip connection is crucial for expanding the recursive multiplicative terms into higher-order interaction terms.

The final output y_t, modeled in Eq. (11), aggregates information from the projected main state (via $\mathbf{C} \in \mathbb{R}^{1\times l}$), the gated state (via $\mathbf{J} \in \mathbb{R}^{1\times m}$), and a direct residual connection $\mathbf{D} \in \mathbb{R}$. Unlike architectures that initialize feedthrough terms to zero, MG-SSM-s explicitly parameterizes $\mathbf{D}$ to preserve immediate signal propagation.

B. Dynamics of the Multiplicative Gate

The recurrence in Eq. (10) expresses non-linearity through sequential multiplication, enabling the model to capture complex dependencies that linear SSMs often miss. By recursively expanding this term, we observe that g_t behaves as a cumulative product interaction through the unfolding steps in Eq. (12) to Eq. (14):

$$g_t = \mathbf{E}\left((\mathbf{E}g_{t-2}) \odot \tilde{x}_{t-1} + \mathbf{F}x_{t-1}\right) \odot \tilde{x}_t + \mathbf{F}x_t \tag{12}$$

$$= \mathbf{E}^2 g_{t-2} \odot (\tilde{x}_{t-1} \odot \tilde{x}_t) + \mathbf{EF}(x_{t-1} \odot \tilde{x}_t) + \mathbf{F}x_t \tag{13}$$

$$= \left(\mathbf{E}^t g_0 \odot \prod_{k=1}^{t} \tilde{x}_k\right) + \sum_{k=1}^{t-1}\left(\mathbf{E}^{t-k}\mathbf{F} \odot \prod_{j=k}^{t} \tilde{x}_j\right) + \mathbf{F}x_t \tag{14}$$

Eq. (14) demonstrates that the skip connection $\mathbf{F}$ transforms the pure product chain into a sum of product interactions of varying lengths. This structure effectively mimics a higher-order interaction mechanism, allowing the network to retain memory of specific input patterns (e.g., $x_{t-1}x_t$) independently of the long-term linear trend, thereby mitigating the risk of vanishing gradients [15].

V. Experiments

A. Dataset

The experimental data for this study was sourced from the publicly available Google COVID-19 Open Data project, specifically utilizing the `table-epidemiology` dataset [16]. This repository is hosted on Google's GitHub platform and provides comprehensive time-series data related to the pandemic. The raw data is distributed as a CSV file, initially covering approximately 231 countries and regions globally.

The complete dataset aggregates multidimensional time-series data, comprising metadata such as timestamps and geospatial locations, alongside a suite of epidemiological indicators. These metrics encompass both daily increments and cumulative totals for confirmed cases, mortalities, recoveries, and testing statistics. From this extensive set of indicators, our research exclusively utilizes the new confirmed cases metric. This variable was prioritized due to its superior data integrity, characterized by robust continuity and a minimal prevalence of missing values across the global cohort compared to other available signals.

To ensure high-quality data sequences for time-series modeling, the initial cohort was filtered down to 40 countries based on demographic and economic scale, as well as absolute data continuity. For the analysis interval spanning from 6th January 2020 to 6th June 2022, countries were excluded if their time series contained any missing values, NaN flags, or anomalous zero-entries resulting from reporting lags or logging omissions.

B. Experiment Setup

1) Environment and dependencies: All experiments were executed within a Google Colaboratory environment utilizing a T4 GPU runtime for computational efficiency. For reproducibility, the primary runtime environment and key library versions used were: Python 3.10; the deep learning framework was PyTorch 2.8.0+cu126; and essential libraries included torchmetrics 1.8.2, statsmodels 0.14.6, Pandas 2.2.2, scikit-learn (sklearn) 1.6.1, and NumPy 2.0.2.

TABLE I. OPTIMAL HYPERPARAMETER SETTINGS FOR COMPARATIVE MODELS.

Model	Latent Dimension (l)	Layers	Specific Configuration
ARIMA	-	-	Tuned (p,d,q) per country
LSTM	256	1	-
Bi-LSTM	256	1	-
GRU	128	1	-
SSM	32	1	-
MG-SSM-s	64	1	Hidden Gate $(m) = 32$

2) Model architectures and configuration: We evaluated six baseline architectures against our proposed MG-SSM-s: ARIMA, LSTM, Bi-LSTM, GRU, and a discrete State Space Model (SSM). The machine learning models were implemented using PyTorch, while the ARIMA model was implemented via the statsmodels library. The SSM baseline adopts the discrete state space formulation of the S4 model, utilizing random initialization rather than the HiPPo matrix to isolate the architectural performance.

To ensure fair comparison, we conducted a grid search for the optimal latent dimension $l \in \{32, 64, 128, 256\}$ for all machine learning models. For the ARIMA models, we conducted a grid search over the parameters $p, q \in \{0, 1, 2, 3, 4, 5\}$ and $d \in \{0, 1, 2\}$. Table I summarizes the optimal hyperparameters identified for each architecture. All models were configured with a univariate input and output size (1×1).

3) Training protocol: A consistent training protocol was applied across all experiments to ensure reproducibility. We utilized a look-back window (input sequence length) of $L = 30$ time steps. Training was performed using the Adam optimizer with a initial learning rate $\alpha = 10^{-3}$ and a batch size of 64. The maximum training duration was set to 1000 epochs, implemented with an early stopping mechanism that halts training if validation loss fails to improve for 50 consecutive epochs.

To facilitate model convergence, all input features were normalized to the range $[-1, 1]$ using Min-Max scaling. This strict bounding is required because the model architecture omits internal normalization layers; constraining inputs to a maximum magnitude of 1 ensures stable activations and prevents exploding gradients. Unbounded alternatives, such as Z-score standardization, were rejected as they cannot guarantee the boundaries necessary to maintain training stability. Finally, to prevent look-ahead bias and data leakage, the data was partitioned sequentially into training, validation, and testing sets using an $8 : 1 : 1$ ratio.

To account for stochastic variations, the pipeline was executed for 21 independent runs ($N_{\text{runs}} = 21$). These runs differed only by their pseudo-random weight initializations and epoch-level batch shuffling order. All model architectures, hyperparameters, and optimization settings were kept strictly identical across all runs.

4) Performance evaluation: The primary metric for comparative evaluation was the Mean Squared Error. The quadratic penalty term in MSE effectively highlights substantial deviations from ground truth, which is critical for ensuring stability in the target application.

TABLE II. GLOBAL PERFORMANCE SUMMARY ACROSS 40 COUNTRIES. RANK DISTRIBUTION DETAILS THE FREQUENCY OF ACHIEVING EACH RANK (1ST TO 6TH) ACROSS THE DATASETS

Model	**Mean Rank**	**Rank Counts**					
		1	2	3	4	5	6
ARIMA	1.975	7	27	6	0	0	0
LSTM	5.725	0	0	0	1	9	30
Bi-LSTM	5.075	0	0	0	6	25	9
GRU	4.10	0	0	2	32	6	0
SSM	2.50	6	8	26	0	0	0
MG-SSM-s	1.625	27	5	6	1	0	1

C. Experimental Results

Table II presents the aggregated performance metrics across the full 40-country cohort. The proposed MG-SSM-s demonstrated remarkable consistency, achieving the first rank in 27 out of 40 datasets and never ranking lower than second. In contrast, ARIMA and the standard Discrete SSM achieved the top rank in seven and six cases, respectively, while the recurrent baselines (LSTM, Bi-LSTM, GRU) rarely achieved competitive rankings.

In the global evaluation across all 40 countries, the average ranking confirms this advantage. MG-SSM-s achieved the best (lowest) mean rank of **1.625**, followed by ARIMA (1.975), Discrete SSM (2.50), GRU (4.10), Bi-LSTM (5.075), and LSTM (5.725).

To validate statistical significance, we performed the Friedman test, which indicated a significant difference among the $k = 6$ models across $N = 40$ datasets ($\chi^2 = 167.228$, $p = 2.846 \times 10^{-34}$).

Subsequent post-hoc analysis using the Nemenyi test ($\alpha = 0.05$) revealed a Critical Difference (CD) threshold of 1.0793. The rank difference between the top three models, MG-SSM-s, ARIMA, and Discrete SSM, was 0.875. Since $0.875 < 1.0793$, the conservative Nemenyi test is insufficient to reject the null hypothesis of equivalence between these three top-performing architectures.

Given this result, we conducted a more sensitive pairwise analysis using the Wilcoxon Signed-Rank Test to directly compare the median MSE distributions of MG-SSM-s against the ARIMA and the Discrete SSM. This test yielded a statistic of $W = 252.0$ with a p-value of $\mathbf{0.016}$. Since $p < 0.05$, we reject the null hypothesis, confirming that MG-SSM-s provides a statistically significant improvement over the Discrete SSM. In contrast, the comparison against ARIMA resulted in $W = 315.0$ and $p = 0.103$. Since $p > 0.05$, we fail to reject the null hypothesis for this pair. These results indicate that while MG-SSM-s significantly outperforms the standard machine learning baselines, its performance is statistically comparable to the tuned ARIMA model.

D. Discussion

The results conclusively establish that the proposed MG-SSM-s is the most robust architecture among the evaluated models. While the initial Nemenyi post-hoc analysis suggested broad statistical overlap between the top-tier models, the pairwise Wilcoxon Signed-Rank Test clarified that the performance gain of MG-SSM-s is statistically significant ($p \approx 0.05$). This

TABLE III. INCREMENTAL ABLATION RESULTS. RANK DISTRIBUTION ACROSS 40 COUNTRIES. THE p-VALUE INDICATES SIGNIFICANCE RELATIVE TO THE PECEDING ROW (PAIRWISE WILCOXON SIGNED-RANK TEST)

Model Variant	Mean Rank	Rank Counts			p-value
		1st	2nd	3rd	
1. Discrete SSM (Baseline)	2.20	14	4	22	-
2. MG-SSM (Gate Only)	2.10	10	16	14	$0.177^{\dagger}$
3. **MG-SSM-s (Gate + Skip)**	**1.70**	**16**	**20**	**4**	$\mathbf{0.010}^{\ddagger}$

† vs. Discrete SSM. ‡ vs. MG-SSM. Metric: Median MSE.

finding aligns with the consistent improvement in mean rank (1.625 vs. 2.50).

We attribute this performance to the architectural design: while the standard Discrete SSM relies on purely linear time-invariant dynamics, the MG-SSM-s introduces a multiplicative gate with an explicit skip connection. This structure enables the capture of the non-linear volatility observed in daily epidemiological data, which the purely linear baseline may over-smooth.

Interestingly, while MG-SSM-s significantly outperformed all neural and state-space baselines, its performance was statistically comparable to the tuned ARIMA model ($p = 0.1032$). This suggests that the linear stochastic components of ARIMA are remarkably effective for the underlying trend of this dataset. However, MG-SSM-s achieved a superior mean rank (1.625 vs. 1.975 for ARIMA).

We attribute this stability to the architectural design: while ARIMA relies on fixed lag observations, the MG-SSM-s uses a multiplicative gate with a skip connection. This allows the model to switch between long-term trends and reacting to the non-linear volatility often found in daily epidemiological data

E. Ablation Study

To validate the necessity of each architectural component, we conducted a step-wise ablation study comparing the full model (**MG-SSM-s**) against two variants: the gate-only model (**MG-SSM**) and the standard discrete baseline (**SSM**).

We omitted the skip-only variant, as removing the multiplicative term reduces the gate equation to a simple linear projection ($g_t = \mathbf{F}x_t$). This would functionally duplicate the main linear state h_t, resulting in a redundant standard SSM with increased parameters rather than a distinct architectural mechanism.

1) Impact of the skip connection: The primary theoretical contribution of this work is the stabilization of the multiplicative gate via the skip connection ($\mathbf{F}$). As derived in Eq. (14), the gate-only variant (MG-SSM) is governed by Eq. (15):

$$\begin{aligned} g_t &= \mathbf{E}g_{t-1} \odot \tilde{x}_t \\ &= \mathbf{E}^t g_0 \odot \left(\prod_{k=1}^{t} \tilde{x}_k \right) \end{aligned} \tag{15}$$

This cumulative product poses a risk of gradient instability. Empirically, this instability was evident in Table III. While the gate-only MG-SSM improved slightly over the baseline, the improvement was not statistically significant ($p = 0.177$), likely due to optimization difficulties in the pure product chain.

Fig. 2. Pairwise comparison of Median MSE for MG-SSM (Gate Only) vs. MG-SSM-s (Full). Points below the diagonal indicate countries where the addition of the skip connection reduced the error.

However, the introduction of the skip connection (Row 3) drastically stabilized performance. MG-SSM-s achieved a statistically significant improvement over the gate-only variant ($p = 0.010$), shifting the rank distribution such that it rarely fell to last place (only 4/40 cases). Visual evidence of this improvement is presented in Fig. 2.

2) Impact of the multiplicative gate: The comparison between the baseline SSM and the MG-SSM (Rows 1 vs. 2) isolated the effect of the multiplicative mechanism. Although the MG-SSM reduced the count of worst-case performances (Rank 3 dropped from 22 to 14), it failed to consistently outperform the linear baseline without the skip stabilizer. This confirmed that while the multiplicative gate provides high-capacity dynamics, the skip connection is a requisite component to harness this capacity effectively.

VI. CONCLUSION

The rapid propagation of the COVID-19 pandemic has placed unprecedented strain on global healthcare systems, particularly in densely populated nations. In this context, accurate forecasting of confirmed case counts is critical, providing policymakers with the necessary intelligence to anticipate surges, allocate resources, and implement effective public health measures. Furthermore, reliable predictive data serves to motivate public adherence to mitigation strategies intended to curb the virus's spread.

In this study, we evaluated the efficacy of five deep learning architectures—LSTM, Bi-LSTM, GRU, SSM, the proposed MG-SSM-s, and ARIMA using the Google COVID-19 Open Data repository. The analysis encompassed daily confirmed cases across 40 distinct countries, including major epicenters such as the USA, India, Brazil, France, and Germany. This methodological approach utilized a 30-day look-back window

to forecast next-day confirmed cases, capitalizing on the inherent capacity of deep learning models to capture non-linear temporal dependencies within complex epidemiological data.

Performance verification based on the Mean Squared Error metric demonstrates that the MG-SSM-s model achieves superior forecasting accuracy compared to all deep learning baselines, while delivering comparable statistical performance to a tuned ARIMA model. These results highlight the potential of the MG-SSM-s framework as a robust tool for epidemiological time-series forecasting.

The results demonstrate that the proposed MG-SSM-s is a highly robust architecture among the evaluated models. While the initial Nemenyi post-hoc analysis suggested statistical overlap among the top-tier models, the pairwise Wilcoxon Signed-Rank Test clarified that the performance gain of MG-SSM-s over the standard SSM is statistically significant ($p \approx 0.05$), aligning with its superior mean rank (1.625 vs. 2.50). Interestingly, while MG-SSM-s significantly outperformed all neural and state-space baselines, its performance was statistically comparable to the tuned ARIMA model (p = 0.1032), though MG-SSM-s maintained a superior mean rank (1.625 vs. 1.975 for ARIMA). This performance and stability stem from the architectural design: while standard SSMs rely on purely linear dynamics and ARIMA relies on fixed lag observations, MG-SSM-s introduces a multiplicative gate with an explicit skip connection. This structure enables the model to dynamically switch between tracking long-term trends and capturing the non-linear volatility observed in daily epidemiological data without over-smoothing.

A. Limitations and Future Work

While the framework demonstrates strong forecasting capabilities, its current constraints highlight several avenues for future research:

- Univariate Input Constraint: The model currently relies solely on historical case numbers. Future iterations will extend the architecture to multivariate inputs to incorporate exogenous variables like government policy changes and mobility indexes.
- Data Evaluation Scope: The empirical evaluation was restricted to 40 countries chosen primarily as major pandemic epicenters. Broader validation across diverse regional data qualities is required to verify global generalizability.
- Look-Back Sensitivity: The analysis relied on a fixed 30-day look-back window. Investigating how varying or adaptive window horizons affect forecasting accuracy remains an open direction for exploration.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the author(s) used Google's Gemini 2.5 Flash, in order to refine and improve the clarity, grammar, and flow of the language in the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published study.

References

[1] H. Hu, K. Nigmatulina, and P. Eckhoff, "The scaling of contact rates with population density for the infectious disease models," *Mathematical biosciences*, vol. 244, no. 2, pp. 125–134, 2013.

[2] O. Postavaru, S. R. Anton, and A. Toma, "Covid-19 pandemic and chaos theory," *Mathematics and Computers in Simulation*, vol. 181, pp. 138–149, 2021.

[3] S. I. Alzahrani, I. A. Aljamaan, and E. A. Al-Fakih, "Forecasting the spread of the covid-19 pandemic in saudi arabia using arima prediction model under current public health interventions," *Journal of infection and public health*, vol. 13, no. 7, pp. 914–919, 2020.

[4] J. Yuan, Y. Wu, W. Jing, J. Liu, M. Du, Y. Wang, and M. Liu, "Non-linear correlation between daily new cases of covid-19 and meteorological factors in 127 countries," *Environmental research*, vol. 193, p. 110521, 2021.

[5] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm," *Chaos, Solitons & Fractals*, vol. 140, p. 110212, 2020.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Y. Kong, Z. Wang, Y. Nie, T. Zhou, S. Zohren, Y. Liang, P. Sun, and Q. Wen, "Unlocking the power of lstm for long term time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 11, 2025, pp. 11 968–11 976.

[8] W. Tangseefa, T. Pumpaibool, P. Khanarsa, and K. Sinapiromsaran, "Forecast covid-19 epidemics by strengthening deep learning models with time series analysis." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 7, 2025.

[9] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[11] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[12] J. T. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022.

[13] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[14] W. L. Brogan, *Modern control theory*. Pearson education india, 1985.

[15] S.-H. Noh, "Analysis of gradient vanishing of rnns and performance comparison," *Information*, vol. 12, no. 11, p. 442, 2021.

[16] G. C. Platform, "Covid-19 open data: Epidemiology table," https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-epidemiology.md, 2020–2022, accessed: 12/11/2025.

Topology-Aware Fingerprint Representation Using Graph Convolutional Network for Robust Minutiae Refinements

Yoogesh A[1], Rama Prasath A[2]*
Research Scholar-Department of Computer Applications
SRM Institute of Science and Technology, Tiruchirappalli, India 621105[1]
Department of Computer Applications, SRM Institute of Science and Technology,
Tiruchirappalli, India 621105[2]

***Abstract*—Most of the fingerprint recognition systems represent minutiae as independent points. This reduces robustness under noise, distortion, and partial impressions. The lack of explicit structural modeling contributes to inconsistent feature reliability, specifically in defocused acquisition conditions. To address these limitations, a Topology-Aware Graph Convolutional Network (Topo-GCN) is proposed. Topo-GCN is a geometric fingerprint representation framework that treats all minutiae points as interrelated nodes and polishes them using a Graph Convolutional Network (GCN). Each node is represented as a 14-dimensional descriptor with the Multi-scale Spatial Feature Tensor (MSFT) technique. A novel Relational Connection Factor (RCF) applies adaptive topology-aware learning to construct edges in place of traditional distance-based graph construction. The Hybrid scoring technique combines learned node representations with graph-theoretic measures to separate genuine minutiae from false ridge information. Furthermore, a lightweight pseudo-labeling scheme efficiently trains the model without depending on the large-scale annotated datasets. The proposed Topo-GCN framework achieves an Area under Curve (AUC) of 0.9614, 0.9769, and 0.9831 with consistent verification performance (EER) of 1.85%, 1.39%, and 1.64% and stable relational encoding with mean edge weights of 0.712, 0.739, and 0.758 across FVC2000, FVC2002, and FVC2004 datasets. The results indicate that integrating relational topology into minutiae modeling considerably enhances robustness, making Topo-GCN a viable approach for secure and efficient fingerprint authentication systems.**

Keywords*—Biometric authentication; fingerprint recognition; Graph Convolutional Network; minutiae refinement; topology-aware representation

I. Introduction

The modern technological era demands high-security systems with a relatively low computational complexity for individual authentication. The inherent vulnerability of password- and token-based authentication mechanisms to security breaches highlights the necessity for developing more robust and reliable alternatives. In this regard, biometric authentication relies on characteristics that are intrinsically associated with an individual, whereas former methods depend on knowledge that can be guessed and imitated.

The computationally inexpensive algorithm requirement narrows down to the utilization of unimodal biometric techniques rather than a multimodal approach that requires heavy processing. Among various traits, fingerprints are superior due to their uniqueness and permanence. Additionally, fingerprint sensors are cost and space-efficient compared to other biometric sensors, as they can be integrated to authentication device with a minimal modification. Similar to traditional techniques, fingerprint-based authentication systems are vulnerable to attacks. Following a successful security breach, unlike password reset, fingerprints cannot be replaced. Therefore, protecting stored fingerprint template against inversion, linkage, and replay attacks is important. Cancelable template approach satisfies the properties of diversity, irreversibility, revocability and unlinkability. Minutiae extraction is one of the keystages in the fingerprint authentication system pipeline.

Existing minutiae extraction approaches rely on handcrafted rules, which often fail to capture the relational structure among minutiae points [1]. They are sensitive to noise and lacks generalization, and may discard valid minutiae. Furthermore, these schemes do not explicitly exploit topological dependencies [2]. The importance of relational and spatial dependency modeling has also been shown in graph-based spatial processing frameworks [3]. These limitations motivate the need for a structurally aware and learnable refinement mechanism.

The major contributions of this work are summarized as follows:

- A hybrid fingerprint processing framework that combines MSFT-driven texture encoding with a learnable graph-based refinement, allowing structurally aware minutiae validation without high computational complexity.
- A novel graph-oriented clipping technique replaces conventional rule-based filtering with adaptive node scoring derived from the feature quality, connectivity pattern, and relational consistency measures.
- Achieving better storage efficiency and improved matching performance by generating a compact and unique representation with selective node retention and the graph embedding technique.

II. Related Work

Techniques related to fingerprint minutiae extraction have advanced considerably. Existing fingerprint recognition sys-

*Corresponding author

tems rely on either one of the following techniques described below. Each technique has a unique way of feature representation. Traditional minutiae extraction approaches primarily rely on Crossing Number (CN) and Gabor filtering, which extract Standard Minutiae Triplets (SMT). These methods suffer from sensor noise, scaling and translation. Recent deep learning-based methods attempt to learn feature representations directly from data, introducing heatmap-based and descriptor-based representations, but ignores intrinsic topological relationships among minutiae. Global representation models rely on fixed-length embeddings rather than conventional technique. Such models are computationally efficient with reduced discriminative capability. Hybrid approaches combine local and global minutiae information to create a balanced structural and statistical representations. However, they still depend on heuristic design choices and does not support adaptive relational modeling. SDK-dependent systems rely on proprietary black-box extractors such as VeriFinger and MegaMatcher. They are limited to reproducibility and end-to-end optimization. The comparative analysis of these approaches is summarized and given below.

The classical minutiae extraction methods include a Gabor filter-based linear band-pass approach utilized to compute local ridge frequency and orientation [1], where Standard Minutiae Triplets (SMT), i.e., spatial coordinates (x, y) and the orientation angle (theta) for ridge endings and bifurcations, are extracted; however, reliance on a centroid-based reference point for coordinate translation is mathematically unstable because the addition of a single false minutia alters the entire coordinate system [1]. The Crossing Number (CN) approach [2] is also utilized, but the Crossing Number (CN) algorithm is incapable of distinguishing between legitimate biological traits and topological artifacts [2].

Similarly, a Gabor filter-based extraction approach [4] depends on a singular global reference point (core point) for the Mobius transformation, which is sensitive to registration jitter, where a one-pixel shift in the reference point causes a total non-linear divergence of the transformed coordinates [4]. The Adaptive Scale-Invariant Feature Transform (SIFT) detector [5] extracts 128-dimensional feature vectors consisting of SMT derived from local gradient histograms; however, SIFT is mathematically designed to locate extrema in continuous intensity gradients, which are destroyed by the binarization and thinning process [5]. The Fingerprint Manual Minutiae Marker (FM3) toolkit [6] utilizes the ISO/IEC 19794-2 triplet, but the use of manual extraction introduces a human subjective error [6].

These paradigms treat minutiae as isolated features, explicitly lacking in learning and modeling the relational topology.

Deep learning-based minutiae extraction methods include a CSPNet-based extractor (adapted Cross Stage Partial Network) trained via knowledge distillation from FingerNet [7], extracting SMT and a confidence level for each minutia; however, the feature extractor limits system performance and mimics the systematic errors and biases due to the reliance on FingerNet knowledge [7]. DeepPrint, a multi-branch deep network that incorporates a minutiae-map detection branch to guide representation learning [8], encodes SMT as a 6-channel heatmap, but discards the local relative topological distances between minutiae, making the system vulnerable to false accepts in high inter-class similarity [8]. A ResNet50-Transformer Minutiae Map Predictor [9], which utilizes a ResNet50 backbone coupled with self-attention transformer layers to predict a 12-channel heat map, extracts SMT and a 96-dimensional descriptor for each minutia patch; however, the reliance on fixed 96×96 patches alters the local texture relative to the minutiae center [9]. Deep learning-based methods improve feature representations by operating on the pixel-level suggestions and failing to encode inter-minutiae relationships.

Global representation models include fixed-length representations [10], where textural cues and minutiae representations are concatenated into a 512-dimensional fixed-length feature vector, but the relative geometric distances between minutiae are ignored [10]. A customized SVM classification model (C-SVM-Cl) [11] maps global P-Hash vectors directly to a biometric identifier (BID), using a 64-bit global hash vector derived from the frequency domain of the fingerprint image, but abandoning local minutiae sacrifices high-entropy discriminative features, resulting in a mediocre 75% accuracy [11]. ShuffleNetv2.3-based architectures [12] extract high-level hierarchical feature vectors as global embeddings; however, the system is spatially variant, lacking rotation and translation invariance and will fail if the input image is not perfectly aligned [12]. Fully minutiae-free approaches utilizing a DINOv2-Base Vision Transformer backbone [13] produce a 1536-dimensional global embedding from the [CLS] token of the final Transformer layer, but lack explicit topological invariance, making the recognition performance highly volatile [13]. Dilated Convolutional Neural Network (DICNN) [14] extracts local and global structural properties such as singularities and ridgeline shapes, but a gridding effect causes sparse spatial sampling that ignores contiguous ridge-valley connectivity [14]. Hierarchical Heterogeneous Ant Colony Optimization based Fingerprint Matching (HHACOFM) [15] relies on ridge topology matching using a boolean matrix representation, but depends entirely on ant agents to compensate for translational and rotational shifts [15]. Pre-trained VGG16 and VGG19 backbones [16] extract high-dimensional global latent embeddings, but lack rotation and translation invariance [16]. A ViT-based Siamese Network [17] produces 256-dimensional representations, but the ResNet-18 front-end discards high-frequency ridge-valley details necessary for sub-pixel localization accuracy [17]. Attention-based and CNN-based classification heads [18] extract global embeddings and local feature maps, but reliance on an external spatial alignment module leads to catastrophic misalignment if inaccurate [18]. Log Gabor feature vectors [19] generate a 477,144-dimensional feature space, but introduce redundancy and increase false positives [19]. CNN-based descriptors utilizing AlexNet [20] produce fixed-size latent feature vectors, but the architecture is incapable of resolving fine-grained ridge-line singularities [20].

The fixed-length global embeddings suppress local geometric relationships. These abstractions discard the reliable discriminative structural information.

Hybrid representations include ResNet50 with multihead attention for global feature extraction [21], extracting SMT, but Sobel-based enhancement cannot distinguish between legitimate ridge boundaries and high-frequency noise artifacts [21]. Minutiae triplet-based feature extraction [22] produces a

6D feature vector including spatial distance, ridge count, and ridge density, but handcrafted topology is sensitive to non-linear skin deformation [22].

Hybrid techniques combine local and global information but still rely on predefined handcrafted features and lacks adaptive modeling.

SDK-dependent systems include COTS SDK (MegaMatcher) [23], extracting SMT of angular sectors, but exhibiting a rotation offset that breaks feature interoperability [23]. VeriFinger SDK 12.0 [24] utilizes paired control points for Thin Plate Spline (TPS) transformation, but errors in the initial transformation cannot be corrected [24]. VeriFinger SDK [25] extracts SMT and core points but is incapable of anchoring focal points in arch patterns [25]. VeriFinger SDK 12.1 [26] used for ground truth extraction prevents gradient-based backpropagation of minutiae-specific loss [26]. Rule-based SDK extraction [27] fails in low-quality regions. VeriFinger 11.1 SDK [28] introduces an opaque dependency preventing optimization. MindTCT with VeriFinger 12.1 [29] constrains the system to classical ridge-following rules. VeriFinger 4.0 SDK [30] fails to distinguish biological singularities from sensor noise. VeriFinger v12.0 [31] introduces circular dependency in degraded regions. Multi-extractor frameworks [32] relying on commercial SDKs create gradient disconnection, preventing optimization of identity loss.

SDK-dependent minutiae extraction methodologies prevent end-to-end optimization and limits the integration of learnable structural representations.

Most state-of-the-art approaches struggle to remain stable with noisy, distorted, and incomplete fingerprint impressions. These limitations necessitate a framework that can capture and adaptively fine-tune the inter-minutiae associations. Topo-GCN models the fingerprint data as a graph, where minutiae are represented as nodes and the structural interactions are encoded through edges, enabling effective learning of relational dependencies.

III. Proposed Work

Despite wide range of advancements, fingerprint-based identification systems validate each minutia independently from the local topology surrounding it. However, ridges are continuous flow fields driven by parallel direction and frequency of directions. Evaluating the isolated minutiae alone creates a structural ambiguity such that spurious detections from noise and broken ridges cannot be categorically separated from real minutiae. Utilizing isolated minutiae to determine the validity of the feature struggles with partial impressions, sensor noise, and elastic distortion. To alleviate this issue, existing methods rely on the enhancement of local descriptors or the learning of global structure of fingerprints. Local descriptors cannot enforce consistency among structures that are spatially distributed, while global embedding models succeed in creating fixed-length representations of fingerprints, downgrading the direct geometric relationship between minutiae. Therefore, both methods provide feature space representations that do not explicitly encode the topological relationship between features, creating instability under occlusion, deformation, and acquisition noise.

A prerequisite to implement meaningful representation of minutiae is to have the ability to establish consistency of relationships amongst minutiae. Therefore, the validity of any given minutia can be determined jointly with neighboring ridge structures. This can be accomplished by retaining both the local features and the underlying topological structure of the fingerprint. Modeling fingerprint minutiae using a graph offers the ideal framework for encoding inter-relational minutiae relationships. However, conventional, naïve graph representations are solely based on spatial proximity and do not reflect the physical continuity of ridge flow, resulting in connectivity problems.

To address these limitations, Topo-GCN framework will define a relational representation that is constrained by ridge-level physical properties and polishes through learnable inference. The proposed approach differs from the traditional approaches by modeling minutiae as a graph and treating it as the primary mechanism of representation. Thus, the validity of a specific minutia arises from structural agreement and not from feature confidence.

In this formulation, the Graph Convolutional Network (GCN) serves an additional purpose beyond merely aggregating features but enforcing consistency across the entire graph. The network iteratively propagates through all of its neighbouring nodes and eliminates the nodes that exhibit structural inconsistency. The utilization of the GCN in the Topo-GCN framework serves as a means of providing distributed relational reasoning via global embedding strategies. The proposed framework shifts minutiae extraction from a detection problem to a consistency-driven inference problem over a structured topological domain.

Fig. 1 illustrates the architecture of the proposed Topo-GCN framework. The extracted fingerprint feature set is represented as a structured graph. Each node encodes a candidate minutia characterized by descriptors derived from the Multi-scale Spatial Feature Tensor (MSFT), including spatial position, orientation with texture, structural, and multi-scale information.

Edges are constructed using the range-restricted connectivity scheme by capturing local ridge topology. The resultant graph is processed with the learnable GCN to adaptively filter node representations by discarding structurally inconsistent candidates. The refinement process improves distinctness by utilizing the learned feature interactions and eliminates the requirement for large-scale training.

A composite scoring mechanism evaluates each node based on a weighted combination of feature confidence, degree centrality, and edge consistency. The nodes with the score exceeding an adaptive threshold are retained, whereas the nodes with low scores are ignored to preserve genuine structural patterns

The proposed scheme inherently supports robustness, compactness, and resistance to direct inversion as the final representation contains relational and topological dependencies rather than raw biometric features.

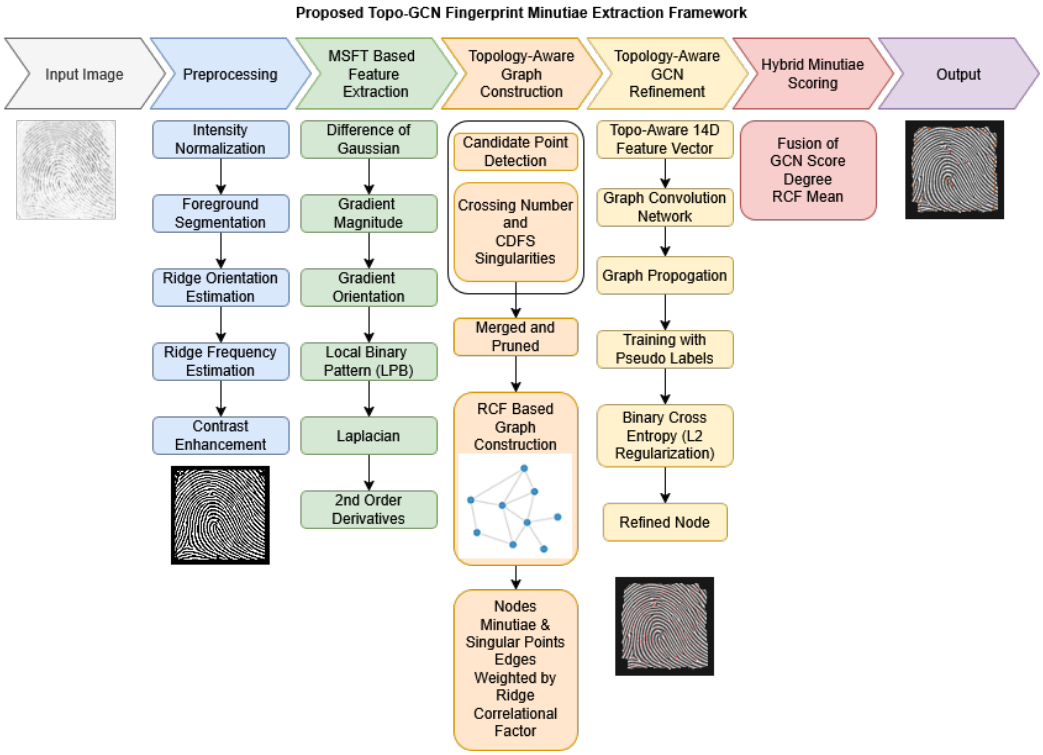


Fig. 1. Proposed fingerprint minutiae extraction framework.

A. Minutiae Representation

The minutiae representation is a three-layered hierarchical construct rather than a single vector. Each detected minutia is represented by a 14-dimensional feature vector X_i used for GCN inference. The feature vector comprises normalized spatial coordinates and orientation (θ/π), local coherence, minutiae type, and mean Ridge Connection Factor (RCF) with seven channels of MSFT to capture ridge-valley contrast. Furthermore, the degree of the node within the RCF graph encodes local structural connectivity. This formulation overcomes the conventional descriptor by redefining minutiae as a relational property instead of an isolated measurement.

The 14-dimensional descriptor represents the minutiae along with the local structural context. This is sufficient for subsequent graph learning.

Unlike DeepPrint, which treats minutiae as parts of a global image map, the proposed system builds a Relational Connection Factor (RCF) graph to capture innate relationships among minutiae. The edge weights f_r are defined as the fusion of Correlation Coefficient r_{cc}, Orientation Consistency o_c, and Frequency Consistency f_c, enabling a physical relationship between minutiae. Ridge flow interruption due to poor sensor contact causes the RCF weight to drop, enabling the GCN to de-weight the error node for better minutiae refinement. The RCF serves as a proxy for ridge-flow continuity. The model extracts structurally consistent minutiae from ridge irregularities, as in Eq. (1):

$$f_r(i,j) = \alpha \cdot r_{cc}(i,j) + \beta \cdot o_c(i,j) + \gamma \cdot f_c(i,j) + \delta \cdot e^{-d_{ij}/\sigma} \quad (1)$$

B. Alignment-Free Property

DeepPrint relies on an "Alignment Module" to orient the fingerprint before feature extraction. Topo-GCN logic is alignment-free and learns the relative topology between nodes. The graph structure remains identical whether the fingerprint is rotated or shifted, provided the relative distances between minutiae are preserved.

The reliance on reference point estimation is eliminated, and the invariance is enforced through relational geometry to make the representation inherently robust to positional variations without explicit normalization.

C. Graph Construction

Instead of Euclidean distance to define graph edges, Topological Relational Construction (TRC) is a parametrically adaptive graph structure that fuses domain-driven constraints with learnable refinement.

The nodes are a fusion of Crossing Number (CN) minutiae and CDFS singularity candidates. Unlike standard pipelines that accept all detected points, the proposed technique prunes them using a coherence threshold and a foreground mask. The pruning criterion is adaptively modulated during training so that the model retains the low-confident yet structurally informative points.

Edge formation is constrained within a local neighborhood. The connectivity is modeled as a continuous function of Euclidean distance D with soft distance-dependent weighting to enable adaptive neighborhood formation.

Each edge is defined through a learnable fusion of structural cues. The edge weight is formulated, as in Eq. (2):

$$f_r(i,j) = W^T \cdot \phi(i,j) \tag{2}$$

where,

$$\phi(i,j) = [r_{cc},\ o_c,\ f_c,\ e^{-d_{ij}/\sigma}]$$

and W is the learnable parameter vector. Initially, the graph is constructed with fixed ridge-based rules and subsequently refined during training through learnable weighting.

D. GCN Adjacency Matrix

For graph convolution, the adjacency matrix is constructed using the refined edge weights and normalized, as in Eq. (3):

$$\hat{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \tag{3}$$

This normalization ensures to distribute relational influence across structurally consistent neighborhoods. Furthermore, the dominance of spurious connections are prevented and true topological interactions are preserved.

E. GCN Propagation

The model propagates through a $14 \rightarrow 32 \rightarrow 16 \rightarrow 1$ architecture in Eq. (4), Eq. (5), and Eq. (6):

$$H_1 = \sigma(\hat{A}X_{\text{norm}}W_1) \tag{4}$$

$$H_2 = \sigma(\hat{A}H_1W_2) \tag{5}$$

$$s_c = \text{Sigmoid}(\hat{A}H_2W_3) \tag{6}$$

The propagation does not end with the GCN output. The final score is obtained, as in Eq. (7):

$$s_{\text{final}} = w^T[s_c,\ d_{\text{norm}},\ \text{RCF}_m] \tag{7}$$

which is the learnable fusion of GCN inference, node centrality, and ridge continuity, enabling adaptive weighting of structural cues.

GCN functions as a mechanism for propagating each node iteratively to validate it against its spatial neighborhood.

F. Global Descriptor Construction

Following graph-based learning, a higher-level 80-dimensional global descriptor is constructed for verification and validation, as in Eq. (8):

$$F = [H^{32}_{\text{ori}},\ H^{8}_{\text{freq}},\ S^{8}_{\text{min}},\ C^{16}_{\text{cyl}},\ S^{8}_{\text{graph}},\ S^{8}_{\text{ridge}}] \in \mathbb{R}^{80} \tag{8}$$

H^{32}_{ori} and H^{8}_{freq} denote the 32-bin orientation and 8-bin ridge frequency, respectively. These capture the overall directional and periodic characteristics of the fingerprint. The first- and second-order statistics of spatial coordinates, orientation, and local coherence are summarized in S^{8}_{min}.

C^{16}_{cyl} captures the local neighborhood structure by evaluating the average orientation distribution within local patches centered at each minutiae. S^{8}_{graph} handles relational consistency by aggregating GCN-derived node quality score, density, and node degree statistics. S^{8}_{ridge} determines ridge-level quality and consistency by incorporating contrast and image quality metrics.

The proposed descriptor sustains connectivity by ensuring that global representation retains the topology of the fingerprint instead of abstracting it and maintaining a compact representation for efficient storage and better processing [33].

G. Learning Strategy

The Topo-GCN framework integrates domain-driven constraints with learnable optimization, thereby avoiding reliance on fixed heuristic formulations. The edge weighting function $f_r(i,j)$ is defined as a fusion of correlation, orientation, frequency consistency, and distance decay, capturing the structural relationships between minutiae. The contribution weights $\alpha, \beta, \gamma, \delta$ govern this combination during graph construction. Although not explicitly learned, their effect is reflected in the learned transformations of node features and adjacency normalization within the GCN refinement stage, eliminating the need for manual tuning.

Furthermore, the structural cues are optimized indirectly with respect to the binary cross-entropy objective through the learning dynamics of the network.

A pseudo-labeling strategy is adopted to enable effective learning without depending on manually labelled datasets. The strategy programmatically classifies candidate minutiae based on structural integrity and connectivity measures. High-coherence and well-connected nodes are treated as positive samples, and low-confidence candidates are treated as negative samples. This facilitates a lightweight supervised training paradigm using mini-batch optimization with binary cross-entropy loss and regularization. To overcome the collapse of label distributions, class imbalance, and enhanced meaningful supervision, percentile-based fall back strategies are applied.

IV. Results and Discussion

Experimental analysis of the proposed Topo-GCN shows that learning over a 14-dimensional feature representation (X) provides better discriminative capability and generalization

performance compared to raw pixel-based modeling by achieving a training accuracy of 99.80% relative to 97.50%, as reflected in verification metrics.

The necessity for collaborative feature learning increases the complexity of raw image-based models during registration and matching the fingerprint. The proposed framework introduces a closely aligned representation directly encoding ridge coherence and alignment position. This shift enables targeted graph-based refinement, reducing the reliance on deep convolutional architectures.

The Topo-GCN framework segregates structural feature encoding from relational refinement by decoupling representational redundancy. This allows the model to emphasize on learning meaningful inter-minutiae dependencies through graph-based processing.

The architectural design ($14 \rightarrow 32 \rightarrow 16 \rightarrow 1$) enables better optimization within a reduced, information-bound space. This is substantiated with a mean processing time of 0.543 seconds and throughput of 1.84 images per second.

The reduced computational complication is due to the structured feature processing, unlike repeated high-dimensional convolutions.

The stable structural encoding and robustness of the representation are supported by the ridge-flow stability metrics, accomplishing a low orientation field error of 0.0126 rad and high coherence with a mean of 0.913. The topology-aware sparse graph with low graph density (0.0215–0.0312) and controlled mean node degree (2.34–3.27) limits unnecessary connectivity between nodes and the potential for over-smoothing via graph convolution. Therefore, node representations remain discriminative by preserving local ridge relationships and leading to improved separation between genuine and impostor score distributions.

Furthermore, the mean RCF edge weight (≈ 0.797) with low variance indicates stable relational encoding across varying fingerprint qualities. This suggests that the edge construction process is not significantly influenced by noise-sensitive features.

Stable relational modeling reinforces edge weighting (mean = 0.797) to maintain local feature interactions to remain discriminative.

Integrating local cylinder descriptors with global 80-dimensional feature embeddings by the hybrid scoring framework strengthens the verification reliability.

Effective separation between genuine (mean ≈ 0.80) and impostor (mean ≈ 0.25) distributions is achieved with the application of $\tanh$-based normalization, which induces a nonlinear redistribution of similarity scores.

The overlap in the decision space and enhanced discriminability are attained with the margin amplification. Jointly, these results indicate that the proposed framework achieves enhanced verification performance through a synchronized blueprint that consolidates graph-structured learning and distribution-aware score normalization.

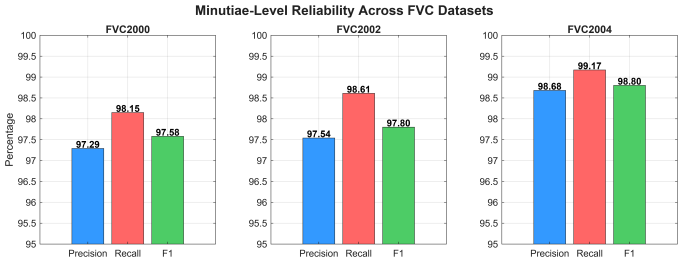


Fig. 2. Visual representation of precision, recall, and F1 score across FVC datasets.

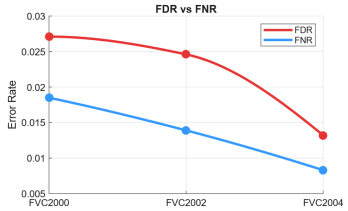


Fig. 3. False Discovery Rate vs. False Negative Rate.

A. Minutiae-Level Reliability (MLR)

To assess the ability of the Topo-GCN framework to filter appropriate candidates, precision, recall, and F1-score are determined. To ensure the successful elimination of noise and retention of genuine ridge features, False Discovery Rate (FDR) and False Negative Rate (FNR) are drawn.

Table I signifies strong minutiae-level reliability across all FVC datasets with F1-scores recorded at 97.58%, 97.80%, and 98.80%, respectively.

As shown in Fig. 2, precision, recall, and F1-score are higher and consistent compared to the state-of-the-art approaches. The higher values of recall suggest greater retention of true minutiae, despite the acquisition conditions.

Moreover, the precision value denotes that the graph-based refinement technique protects true ridge patterns and serves as a strong baseline compared to [14].

Based on the data in Table II, both False Detection Rate (FDR) and False Negative Rate (FNR) decreases from 0.0132 to 0.0083 across datasets denotes the stability of the refinement process.

The proposed framework maintains a strong balance by reducing false detections and preserving true minutiae. The proposed model adjusts to varying quality datasets and does not overfit.

Fig. 3 shows stable refinement behavior. Furthermore, the structural graph-based refinement filters the candidates based on relational consistency rather than isolated feature confidence is one among the improvements over the numerical baseline methods based on trade-off between FDR and FNR.

Table III indicates that the range growing from 2.50 to 3.01, the statistical distance between genuine and impostor score distributions.

Table IV shows that the proposed method achieves lower

TABLE I. MINUTIAE-LEVEL RELIABILITY ACROSS FVC DATASETS

Method	FVC2000			FVC2002			FVC2004		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
[2]	97.13	96.84	96.99	96.29	95.74	96.01	96.5	95.8	96.1
[7]	86.76	89.58	88.51	90.76	91.42	91.09	84.8	86.6	85.7
[14]	96.50	97.69	97.09	95.81	94.96	95.38	98.63	98.47	98.55
[16]	94.42	91.36	92.86	93.64	92.88	93.26	92.6	92.4	92.4
Proposed	97.29	98.15	97.58	97.54	98.61	97.80	98.68	99.17	98.80

TABLE II. FALSE DETECTION RATE (FDR) AND FALSE NON-MATCH RATE (FNR) ACROSS DATASETS.

Dataset	FDR	FNR
FVC2000	0.0271	0.0185
FVC2002	0.0246	0.0139
FVC2004	0.0132	0.0083

TABLE III. SEPARABILITY INDEX (d')

Metric	FVC2000	FVC2002	FVC2004
d' separability	2.50	2.83	3.01

EER across all datasets, outperforming existing methods, particularly on FVC2002 with 1.39%.

Table V further confirms the discriminative capability, achieving AUC values of 0.9614, 0.9769, and 0.9831 for the FVC datasets, respectively.

Fig. 4 shows strong ROC performance, with consistent curves across datasets.

A consistent improvement with the shift in score distributions is observed. The genuine scores cluster around 0.80 while impostor scores remain near 0.25. The hybrid scoring mechanism reinforces the separation by fusing GCN outputs with node degree and ridge continuity. The tanh-based normalization minimizes overlap in the decision space by amplifying decision margins through a non-linear redistribution of similarity scores.

The preprocessing stage influences the graph modeling. Table VI populated with high PSNR and SSIM, significant improvement in CNR, and lower MSE and OFE. This illustrates the necessity of conserving structural information of the image prior to feature extraction [34].

Fig. 5 graphically illustrates the preprocessing and minutiae extraction pipeline. Fig. 6, indicate that the proposed Topo-GCN has improved the quality of the image with stable ridge flow.

Table VII showcases that the low graph density (0.0215–0.0312) and controlled node degree (2.34–3.27) prevent over-smoothing during graph convolution. The existence of clear separation between the GCN score distributions of genuine versus imposter supports the hypothesis that the learned representations of nodes accurately reflect the unique topology of the associated minutiae as opposed to only using local feature descriptors. Therefore, explicit relational modelling improves fingerprint discriminability.

To evaluate RCF edge-weight formulation as in Eq. (1), each component was removed independently to understand the contribution of each term. Table VIII presents the results of the evaluation. Removal of any one of the terms lead to a sharp increase in the EER. The exclusion of r_{cc} recorded the highest EER. The complete formulation of RCF achieved the minimal EER.

TABLE IV. COMPARISON OF EQUAL ERROR RATE (EER)

Models	FVC2000	FVC2002	FVC2004
[5]	2.96	2.36	2.01
[6]	1.38	2.47	1.85
[13]	2.02	2.18	3.48
[17]	2.07	1.61	1.73
[21]	1.85	1.95	2.20
Proposed	1.05	1.39	1.64

TABLE V. AREA UNDER THE CURVE (AUC) ACROSS FVC DATASETS

FVC2000	FVC2002	FVC2004
0.9614	0.9769	0.9831

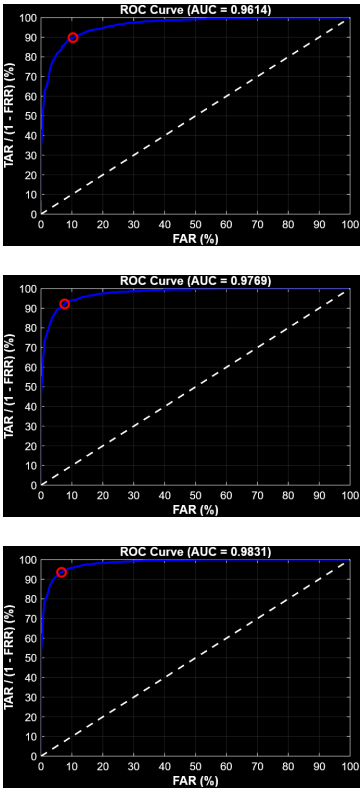


Fig. 4. Receiver Operating Characteristic (ROC) curves of the proposed Topology-Aware Graph Convolutional Network (Topo-GCN): (**a**) FVC2000; (**b**) FVC2002; (**c**) FVC2004.

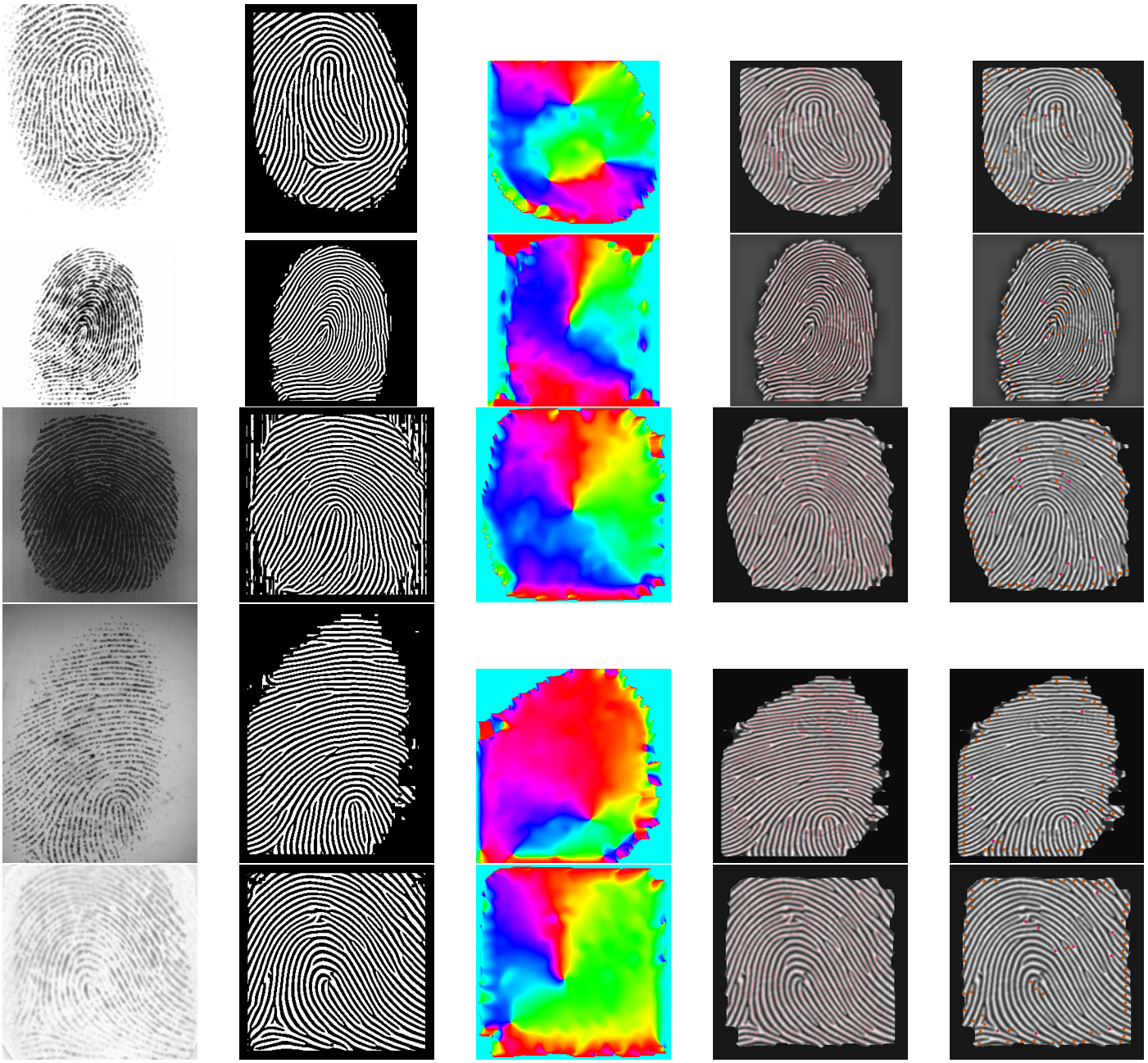


Fig. 5. Qualitative visualization of fingerprint preprocessing and minutiae extraction. Each row corresponds to a different fingerprint sample. Columns (left to right) represent: input image, enhanced image, orientation field estimation, skeletonized ridge structure, and extracted minutiae points.

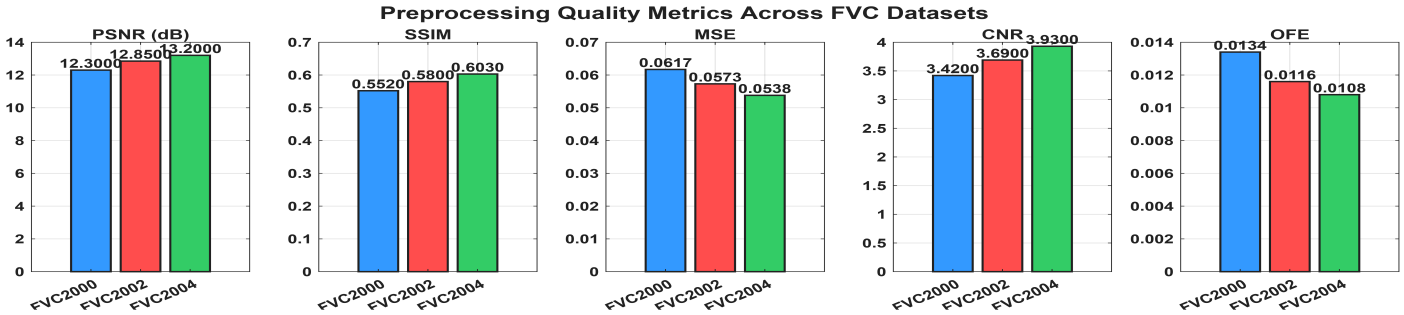


Fig. 6. Graphical representation of preprocessing metrics.

TABLE VI. PREPROCESSING QUALITY METRICS ACROSS FVC DATASETS

Metric	FVC2000	FVC2002	FVC2004
PSNR (dB)	12.30	12.85	13.20
SSIM	0.552	0.580	0.603
MSE	0.0617	0.0573	0.0538
CNR	3.42	3.69	3.93
OFE	0.0134	0.0116	0.0108

Abbreviations: PSNR – Peak Signal-to-Noise Ratio; SSIM – Structural Similarity Index Measure; MSE – Mean Squared Error; CNR – Contrast-to-Noise Ratio; OFE – Orientation Field Error.

TABLE VII. GRAPH-BASED STRUCTURAL METRICS ACROSS FVC DATASETS

Metric	FVC2000	FVC2002	FVC2004
Graph Density	0.0215	0.0268	0.0312
Mean Node Degree	2.34	2.89	3.27
GCN Score (Genuine)	0.742	0.781	0.812
GCN Score (Impostor)	0.284	0.241	0.198
RCF Edge Weight (Mean)	0.712	0.736	0.758
RCF Edge Weight (Std)	0.092	0.085	0.079

TABLE VIII. SENSITIVITY ANALYSIS OF RCF EDGE COMPONENTS

Edge Weight Configuration	EER (%)
Without Correlation Consistency (r_{cc})	2.56
Without Orientation Consistency (o_c)	2.14
Without Frequency Consistency (f_c)	1.83
Full RCF (r_{cc}+o_c+f_c)	1.64

TABLE IX. ABLATION STUDY OF TOPO-GCN FRAMEWORK

Configuration	Precision (%)	Recall (%)	F1 (%)	EER (%)	AUC
Full Model (MSFT + GCN + Hybrid Scoring)	98.68	99.17	98.92	1.64	0.983
Without MSFT Features	96.91	97.42	97.16	2.87	0.964
Without GCN	95.84	96.73	96.28	3.95	0.948
Without Hybrid Scoring	97.12	97.89	97.50	2.41	0.971
Without RCF Graph	94.76	95.88	95.32	4.62	0.932
Without MSFT + GCN	97.45	98.02	97.73	2.18	0.975
Only MSFT	93.28	94.67	93.97	5.84	0.910

B. Evaluation Protocol

Evaluation is carried out under a subject-exclusive protocol, where DB1_B and DB2_B are used for model learning, and DB3_B along with DB4_B are reserved strictly for testing. This partition ensures that no overlap exists between the training and test sets, thus eliminating possible leakages and establishing a realistic scenario for evaluating generalization.

The partitioning method will also provide several differences in the conditions under which samples were obtained within the sub-databases. This permits evaluating a model across distributional differences across FVC Datasets.

The reliability of the performance is derived from the samples in the test by examining the corresponding distributions for genuine and impostor scores. Based on these distributions for the respective test sets, standard performance measures such as AUC (Area Under the Curve) and EER (Equal Error Rate) are calculated to quantify the performance. To assess consistency of these scores across trial instances, confidence intervals are developed using Bootstrap resampling.

Metrics related to minutiae- and verification-level are reported to maintain consistency in the evaluation process.

C. Ablation Study

To assess the performance through ablation analysis, core distinct modules are selectively removed. This experiment confirms that the combination of MSFT-based feature encoding, GCN-based refinement, and hybrid scoring contributes to the effectiveness of the overall system. Therefore, the proposed framework is not a heuristic accumulation but a structurally grounded design.

Additionally, the proposed framework avoids the large-scale labeled data to maintain computational efficiency and generalization capability.

Results of the ablation analysis from Table IX indicate that each core component projects a systematic role in the overall performance of the entire system. The complete model with all cores performs with an F1-score of 98.92%, an EER of 1.64%, and an AUC of 0.983. The removal of the graph structure from the system, decreased the performance with EER = 4.62%, AUC = 0.932, and failed to model the interdependencies between minutiae. These interdependencies are the primary reason for the increase in performance. The removal of the GCN from the system increases the EER to 3.95%, which indicates that the model builds a graph with a larger number of structurally inconsistent candidates. The exclusion of the MSFT features leads to a moderate drop in the quality of the local ridge representation. The addition of the hybrid scoring mechanism further improves the ability to discriminate between candidates, as indicated by an increase in the EER to 2.41% upon its removal. Overall, the results of the analysis show that the sequential degradation of the performance for each configuration confirms that the system has benefited from the complementary effects of feature encoding, graph-based refinement, and adaptive scoring.

V. Conclusion

The proposed Topo-GCN redefines minutiae modeling by explicitly incorporating relational topology through a graph-based fingerprint representation framework. Graph convolutional learning transforms candidate minutiae into a structured graph, effectively capturing the local and global dependencies that are overlooked in conventional point-based representations.

The integration of a compact 14-dimensional feature descriptor with the proposed RCF enables adaptive and physically meaningful edge construction. The hybrid scoring mechanism ensures robust refinement of minutiae under noisy and partial conditions. The adoption of pseudo-labeling further enhances the practicality of the framework by reducing dependence on large annotated datasets. Experimental results across standard FVC benchmarks demonstrate consistent improvements in discriminability and efficiency, validating the effectiveness of topology-driven learning for biometric representation. Unlike deep image-based models that rely on high-dimensional feature extraction, the proposed method achieves competitive performance within a low-dimensional and computationally efficient framework.

Future work will focus on strengthening cross-dataset generalization, incorporating adversarial robustness analysis, and extending the framework toward fully cancelable biometric template generation with formal security guarantees. The findings of this work highlight the importance of structurally-aware representations and establish a foundation for next-generation fingerprint authentication systems.

References

[1] M. Imran, M. S. Umar, and F. Ahmad, "A non-invertible secure template generation using AES encrypted MCC and random triangle projection," *IEEE Access*, vol. 13, pp. 78194–78213, 2025.

[2] S. P. Singh, D. K. Nishad, and S. Khalid, "Enhancing fingerprint identification using fuzzy-ANN minutiae matching," *Measurement: Sensors*, vol. 37, p. 101809, 2025.

[3] C. P. Shahina and B. Mohammed Ismail, "G-Tree indexing and ACO for spatial query," in *Proc. 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, 2021, pp. 200–203.

[4] M. Imran, M. S. Umar, and S. Malhotra, "Privacy preserving cancellable template generation for crypto-biometric authentication system," *IEEE Access*, vol. 13, pp. 158322–158339, 2025.

[5] S. Bakheet, S. Alsubai, A. Alqahtani, and A. Binbusayyis, "Robust fingerprint minutiae extraction and matching based on improved SIFT features," *Applied Sciences*, vol. 12, no. 12, p. 6122, 2022.

[6] D. Sadhya, "Achieving unlinkability in fingerprint templates via k-anonymity and random projection," *Sadhanā*, vol. 49, no. 3, p. 224, 2024.

[7] Z. Jia and C. Huang, "Automated framework for extracting and restoring minutiae from low-quality fingerprints," *IEEE Signal Processing Letters*, vol. 32, pp. 2194–2198, 2025.

[8] J. J. Engelsma, K. Cao, and A. K. Jain, "Learning a fixed-length fingerprint representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1981–1997, 2021.

[9] S. A. Grosz and A. K. Jain, "Latent fingerprint recognition: Fusion of local and global embeddings," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5691–5705, 2023.

[10] T. Kim, D. Kwon, A. K. Das, and Y. Park, "Robust and lightweight user authentication scheme using deep learning-based cancelable biometrics in smart home environments," *IEEE Internet Things J.*, 2026.

[11] C. Guo *et al.*, "A novel biometric authentication scheme with privacy protection based on SVM and ZKP," *Computers & Security*, vol. 144, p. 103995, 2024.

[12] M. Ragab *et al.*, "Enhancing cybersecurity in higher education institutions using optimal deep learning-based biometric verification," *Alexandria Eng. J.*, vol. 117, pp. 340–351, 2025.

[13] B. Arslan, "Minutiae-free fingerprint recognition via vision transformers: An explainable approach," *Applied Sciences*, vol. 16, no. 2, p. 1009, 2026.

[14] N. S. Babu, M. Elamparithi, and V. Anuratha, "An intelligent system for fingerprint recognition and verification using dilated convolutional neural network and extreme learning machine," *Int. J. Membrane Sci. Technol.*, vol. 10, no. 2, pp. 2707–2715, 2023.

[15] N. K. Sreeja, "A hierarchical heterogeneous ant colony optimization based fingerprint recognition system," *Intell. Syst. Appl.*, vol. 17, p. 200180, 2023.

[16] R. Garg *et al.*, "Fingerprint recognition using convolution neural network with inversion and augmented techniques," *Syst. Soft Comput.*, vol. 6, p. 200106, 2024.

[17] Y. Qiu, H. Chen, X. Dong, Z. Lin, I. Y. Liao, M. Tistarelli, and Z. Jin, "IFViT: Interpretable fixed-length representation for fingerprint matching via vision transformer," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 559–573, 2025.

[18] S. A. Grosz and A. K. Jain, "AFR-Net: Attention-driven fingerprint recognition network," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 1, pp. 30–42, 2024.

[19] RSP, T. Thomas, and S. Emmanuel, "Cancelable biometric template generation using random feature vector transformations," *IEEE Access*, vol. 12, pp. 32064–32079, 2024.

[20] T. Gernot and C. Rosenberger, "Robust biometric scheme against replay attacks using one-time biometric templates," *Computers & Security*, vol. 137, p. 103586, 2024.

[21] T. Sawhney *et al.*, "Fingerprint matching for noisy and distorted patterns using a Siamese network with ResNet50 and multihead attention," *IEEE Access*, vol. 13, pp. 88047–88064, 2025.

[22] S. Salma *et al.*, "Optimizing fingerprint identification: CNNs with raw images versus handcrafted features for real-time systems," *IEEE Access*, vol. 13, pp. 108192–108211, 2025.

[23] A. N. Bonacim *et al.*, "Machine learning for fingerprint ridge counting," *IEEE Trans. Biometrics Behav. Identity Sci.*, vol. 8, no. 1, pp. 60–70, 2026.

[24] X. Guan, J. Feng, and J. Zhou, "Phase-aggregated dual-branch network for efficient fingerprint dense registration," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 5712–5724, 2024.

[25] I. Kavati *et al.*, "Design of a fingerprint template protection scheme using elliptical structures," *ICT Express*, vol. 7, no. 4, pp. 497–500, 2021.

[26] A. Wahab *et al.*, "Latent fingerprint enhancement for accurate minutiae detection," *Procedia Comput. Sci.*, vol. 246, pp. 1558–1567, 2024.

[27] S. M. Abdullahi *et al.*, "Cancelable fingerprint template construction using vector permutation and shift-ordering," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 5, pp. 3828–3844, 2023.

[28] S. M. Abdullahi, H. Wang, and T. Li, "Fractal coding-based robust and alignment-free fingerprint image hashing," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2587–2601, 2020.

[29] X. Yin *et al.*, "A novel length-flexible lightweight cancelable fingerprint template for privacy-presing authentication systems in resource-constrained IoT applications," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 877–892, 2023.

[30] W. Yang *et al.*, "Biometrics based privacy-preserving authentication and mobile template protection," *Wireless Commun. Mobile Comput.*, vol. 2018, p. 7107295, 2018.

[31] Z. Pan, Y. Duan, X. Guan, J. Feng, and J. Zhou, "Latent fingerprint matching via dense minutia descriptor," in *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–10.

[32] K. P. Wijewardena, S. A. Grosz, K. Cao, and A. K. Jain, "Fingerprint template invertibility: Minutiae vs. deep templates," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 744–757, 2023.

[33] B. Mohammed Ismail, T. Bhaskara Reddy, and B. Eswara Reddy, "Spiral architecture based hybrid fractal image compression," in *Proc. International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, 2016, pp. 21–26.

[34] Z. Ullah, L. Qi, D. Binu, B. R. Rajakumar, and B. M. Ismail, "2-D canonical correlation analysis based image super-resolution scheme for facial emotion recognition," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 13911–13934, 2022.

Architecting a Low-Latency RAG System for Fast-Moving Consumer Goods (FMCG) Customer Support: A Case Study in Industrial Software Deployment

Meredita Susanty, Alghifari Rasyid Zola
Ilmu Komputer, Fakultas Sains dan Ilmu Komputer, Universitas Pertamina, Indonesia

***Abstract*—Deploying large language models (LLMs) in industrial customer support environments require balancing response accuracy with system latency. This study presents the software architecture and implementation of a Retrieval-Augmented Generation (RAG) system designed for the Fast-Moving Consumer Goods (FMCG) sector. Addressing the limitations of generic LLMs in domain-specific knowledge tasks, we engineered a retrieval-augmented inference pipeline that integrates unstructured data ingestion, Pinecone for vector indexing, and Groq-based inference for low-latency response generation. The proposed system aims to improve response grounding by incorporating organizational product information into the generation process while maintaining responsive interaction times suitable for customer support applications. This study details the software architecture, system integration approach, and experimental evaluation of the proposed deployment-oriented RAG framework in an industrial FMCG case-study setting.**

***Keywords*—Software architecture; RAG; industrial AI; Groq LPU; vector database; customer support systems**

I. Introduction

Large language models (LLMs) have become increasingly popular in customer service applications due to their ability to generate natural language responses and assist users across a wide range of tasks [1]. Organizations are progressively exploring LLM-based solutions to improve service quality, reduce operational workload, and provide faster access to information [1], [2]. In the Fast-Moving Consumer Goods (FMCG) sector, customer support teams frequently handle inquiries related to product specifications, ingredients, usage instructions, and company procedures. As product portfolios evolve rapidly and organizational knowledge changes continuously, maintaining accurate and consistent responses across customer support channels becomes a significant challenge [2].

Despite their capabilities, general-purpose LLMs exhibit fundamental limitations when applied to enterprise-specific domains. These models primarily rely on parametric knowledge acquired during pre-training and therefore lack direct access to proprietary, confidential, or recently updated organizational information [3]. Consequently, when queried about domain-specific products, internal procedures, or newly released information, LLMs may generate plausible yet factually incorrect responses, a phenomenon commonly referred to as hallucination [4]. In customer support settings, such inaccuracies may negatively affect customer trust, service consistency, and operational effectiveness [5].

Retrieval-Augmented Generation (RAG) has emerged as a promising approach for mitigating these limitations by integrating external knowledge sources into the response generation process [6]. Instead of relying solely on internal model parameters, RAG systems retrieve relevant information from an external knowledge repository and provide this information as contextual evidence for generation [6]. This architecture enables organizations to update knowledge independently of the language model itself, making RAG particularly suitable for dynamic business environments where information changes frequently [7].

Although RAG can improve factual grounding, the architecture introduces additional computational stages, including query processing, retrieval, context assembly, and response generation [2]. These stages increase overall response latency compared to direct LLM inference. In customer support scenarios, where users typically expect responses within only a few seconds, excessive latency may reduce perceived responsiveness and negatively impact user experience [2]. Therefore, practical deployment of RAG systems requires balancing response quality with operational performance [7].

Recent advances in inference infrastructure have created opportunities to improve the responsiveness of Retrieval-Augmented Generation systems. Among these developments, Groq's inference platform has attracted attention for its ability to provide low-latency language model inference and high token-generation throughput. While previous studies have extensively investigated retrieval effectiveness and answer quality in RAG systems [6], comparatively less attention has been given to deployment-oriented architectures that simultaneously consider retrieval-based grounding and low-latency response requirements for enterprise customer support applications [7].

This study presents the design and evaluation of a low-latency RAG architecture for FMCG customer support. The proposed system integrates document ingestion, vector-based knowledge retrieval using Pinecone, and LLM inference served through Groq's infrastructure. Rather than proposing a novel retrieval algorithm, this work focuses on the architectural design and practical implementation of a retrieval-augmented customer support system using real industrial knowledge assets [7]. The objective is to investigate whether retrieval-based grounding can improve response accuracy while maintaining latency suitable for interactive customer support scenarios [2].

The main contributions of this study are summarized as follows:

- The design and implementation of a Retrieval-Augmented Generation architecture for FMCG customer support applications.
- An empirical evaluation of response accuracy and end-to-end latency using domain-specific customer support queries and proprietary organizational knowledge.
- An empirical comparison between a local GPU-based deployment and a Groq-based deployment within a retrieval-augmented customer support pipeline.
- A discussion of practical implementation considerations, limitations, and lessons learned from developing a deployment-oriented RAG system for enterprise use cases.

II. Related Work

A. Fine-Tuning and Retrieval-Augmented Generation

Several approaches have been proposed to adapt large language models (LLMs) to domain-specific applications [3]. One widely adopted approach is fine-tuning, where pre-trained models are further trained on task-specific or domain-specific datasets to improve performance in specialized contexts. Previous studies have demonstrated that fine-tuning can improve downstream task performance by incorporating domain knowledge into model parameters [3].

However, fine-tuning presents several challenges in enterprise environments. Updating organizational knowledge typically requires additional training cycles, computational resources, and model maintenance [7]. As organizational information evolves over time, maintaining an up-to-date fine-tuned model can become costly and operationally complex. Furthermore, fine-tuned models may continue to rely on outdated information embedded within model parameters when knowledge sources change frequently [7].

To address these limitations, Retrieval-Augmented Generation (RAG) was introduced as an alternative architecture that separates knowledge storage from language generation [6]. Instead of embedding all knowledge within model parameters, RAG retrieves relevant information from an external knowledge source and incorporates the retrieved context into the generation process [6]. Similar findings were reported by Guu et al. [8], who demonstrated that retrieval-enhanced architectures can improve factual grounding while maintaining the reasoning capabilities of large language models.

Recent studies have further highlighted the advantages of retrieval-based systems for knowledge-intensive applications [6]. Kandpal et al. [9] observed that large language models often struggle to accurately recall long-tail or domain-specific information, reinforcing the importance of external retrieval mechanisms for enterprise knowledge management scenarios.

B. Retrieval-Augmented Generation in Enterprise Knowledge Systems

The adoption of Retrieval-Augmented Generation has expanded beyond academic benchmarks into practical enterprise applications [6]. Organizations increasingly utilize RAG architectures for knowledge management, document question answering, internal search systems, and customer support automation [2]. By retrieving relevant documents from external repositories, RAG systems can provide responses grounded in organizational knowledge while reducing the likelihood of hallucinated content [6].

Gao et al. [7] identified document chunking, retrieval quality, and context construction as critical components influencing overall RAG performance. Similarly, several enterprise-oriented implementations have demonstrated the effectiveness of vector databases and retrieval pipelines in improving access to large collections of unstructured organizational information [5].

Despite these advances, most prior work focuses primarily on retrieval effectiveness and answer quality [6], [7]. Comparatively fewer studies investigate deployment-oriented considerations such as inference latency, architectural integration, and operational constraints encountered in enterprise environments [2]. As a result, understanding how retrieval-enhanced systems can be implemented while maintaining acceptable response times remains an important area of research.

C. Low-Latency LLM Inference Architectures

While retrieval can improve factual grounding, additional processing stages introduce latency into the overall system [6]. Consequently, recent research has increasingly focused on optimizing large language model inference to support real-time applications [10].

Traditional GPU-based deployments remain the dominant approach for serving transformer-based language models. To improve inference efficiency, several optimized serving frameworks have been proposed, including vLLM and TensorRT-LLM. These frameworks improve throughput and memory utilization through techniques such as paged attention, kernel optimization, efficient batching strategies, and optimized execution pipelines [10].

In parallel, specialized inference hardware has also emerged as an alternative approach for reducing inference latency. One notable example is Groq's tensor streaming architecture, which employs deterministic execution and compiler-driven optimization to improve token generation throughput and reduce inference overhead. Such approaches aim to improve responsiveness for interactive AI applications where low latency is a critical requirement [10].

Although these approaches significantly improve inference performance, existing studies primarily evaluate hardware acceleration or serving optimizations independently from retrieval-enhanced architectures. Limited research has examined how low-latency inference infrastructures interact with end-to-end Retrieval-Augmented Generation pipelines in enterprise customer support settings [6], [2].

D. Research Gap

Based on the reviewed literature, prior studies have extensively investigated the benefits of retrieval augmentation for improving factual accuracy and reducing hallucinations. Likewise, substantial efforts have been devoted to accelerating

large language model inference through specialized hardware and optimized serving frameworks.

However, limited attention has been given to deployment-oriented architectures that simultaneously address retrieval-based knowledge grounding and low-latency response requirements in enterprise customer support environments. Furthermore, relatively few studies report practical implementation experiences using real organizational knowledge assets within industrial case-study settings.

Therefore, this study focuses on the design and evaluation of a low-latency Retrieval-Augmented Generation architecture for FMCG customer support. Rather than proposing a new retrieval algorithm or language model, the work investigates how retrieval-based grounding and different deployment configurations can be combined within a practical enterprise-oriented architecture to improve response quality while maintaining interactive response times.

III. Methods

This study adopts a Retrieval-Augmented Generation (RAG) architecture designed to support domain-specific question answering within FMCG customer support scenarios. The proposed system combines document preprocessing, vector-based retrieval, prompt augmentation, and large language model inference into a unified pipeline. The methodology consists of four primary stages: knowledge base construction, retrieval pipeline development, response generation, and performance evaluation.

A. System Overview

The overall architecture of the proposed system is illustrated in Fig. 1. The system follows a modular design consisting of four primary components: document ingestion, vector retrieval, context augmentation, and language model inference.

During the offline phase, organizational documents are processed and transformed into retrieval-ready chunks. These chunks are subsequently converted into vector representations and indexed within a Pinecone vector database. During the online inference phase, user queries are transformed into vector representations using the same procedure and submitted to the retrieval system. The most relevant document chunks are retrieved and incorporated into a structured prompt before being passed to the language model for response generation.

This architecture separates knowledge storage from reasoning, allowing organizational knowledge to be updated without modifying the underlying language model.

B. Knowledge Base Construction and Chunking Strategy

The knowledge base used in this study consists of proprietary FMCG product information documents from one brand within a portfolio of fourteen FMCG brands. The selected brand contains a comprehensive collection of product specifications, ingredients, product descriptions, and usage-related information commonly referenced during customer support interactions. This single-brand repository was used as an industrial case-study dataset to evaluate the proposed Retrieval-Augmented Generation architecture in a realistic customer support scenario.

Prior to indexing, all documents underwent a preprocessing stage to generate retrieval-ready chunks. As illustrated in Fig. 2, documents were initially segmented into paragraphs, after which non-informative segments and excessively short paragraphs were removed. Longer paragraphs were subsequently divided into smaller sentence groups to improve retrieval granularity while preserving local contextual coherence. This chunking strategy was designed to reduce context fragmentation and increase the likelihood that retrieved chunks contain complete product-related information. Each resulting chunk was associated with relevant metadata and prepared for vector indexing within the retrieval pipeline.

C. Vector Representation and Retrieval Pipeline

Following document segmentation, each chunk was transformed into a fixed-length vector representation to support similarity-based retrieval. The current implementation employs a lightweight hash-based vectorization approach rather than a transformer-based sentence embedding model. Individual lexical tokens are mapped into a 1024-dimensional vector space using a deterministic hashing function, producing a fixed-length representation for each document chunk. The resulting vectors are subsequently normalized prior to similarity comparison. This approach was selected due to its implementation simplicity and compatibility with the lightweight retrieval architecture adopted in this study. This choice was made to align with the lightweight prototype implementation, where vectorization was performed directly within the application service without requiring a separate neural embedding inference component. Consequently, the retrieval layer could be deployed with minimal additional infrastructure and computational overhead.

Because the adopted representation is primarily based on lexical features, it captures token-level similarity rather than the semantic relationships typically learned by neural embedding models. Consequently, retrieval performance may be affected when semantically similar queries exhibit substantially different lexical forms. The generated vectors are indexed using Pinecone as the retrieval backend. During inference, incoming user queries undergo the same vectorization procedure before similarity search is performed against the indexed knowledge base. For the experiments reported in this study, the retrieval component returns the Top-K most relevant document chunks, where $K = 15$. These retrieved chunks are subsequently incorporated into the prompt construction stage and supplied to the language model as contextual evidence for response generation.

D. Retrieval-Augmented Response Generation

After retrieval, the selected document chunks are assembled into a structured prompt and supplied to the language model for response generation. The retrieved information serves as contextual evidence, allowing the model to generate responses grounded in the available product knowledge rather than relying solely on its internal parametric knowledge.

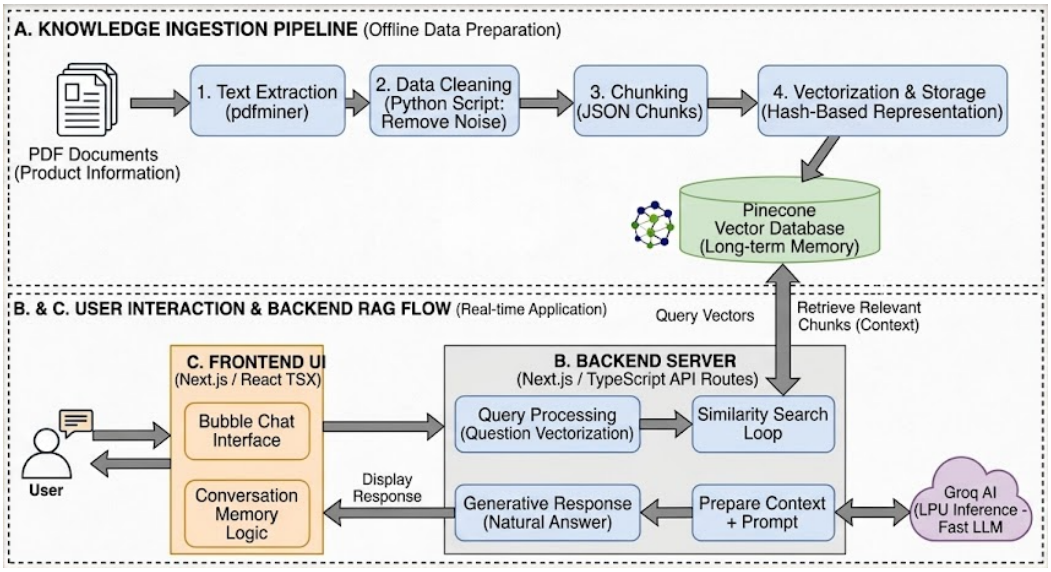


Fig. 1. Overall architecture of the proposed Retrieval-Augmented Generation pipeline.

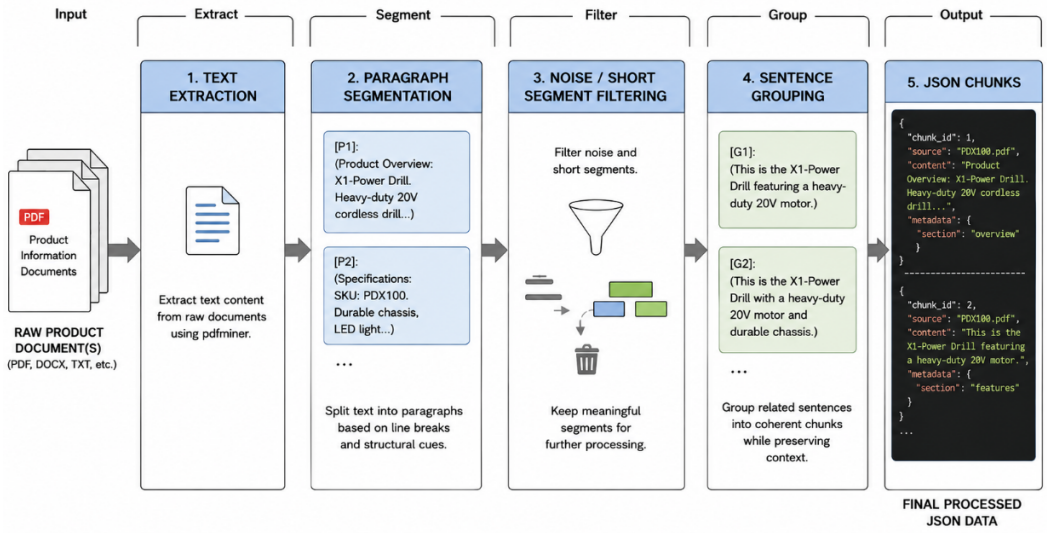


Fig. 2. Knowledge base construction and chunking strategy used for document preparation.

To improve factual consistency, the system adopts a closed-context prompting strategy. As illustrated in Fig. 3, the prompt explicitly instructs the model to generate responses only from the retrieved information and to avoid introducing unsupported content beyond the available context. This approach aims to reduce hallucinations by constraining the generation process to the evidence provided by the retrieval stage.

The overall inference workflow is illustrated in Fig. 4. Upon receiving a user query, the system performs retrieval, prompt construction, and response generation sequentially before returning the final response to the user. For language generation, this study utilizes the Llama-3.3-70B-Versatile model served through Groq's inference infrastructure. A temperature value of 0 was selected to improve output consistency and reduce response variability, while the maximum output length was limited to 600 tokens.

Groq's inference platform was selected due to its low-latency serving capabilities and compatibility with interactive question-answering applications. The objective of incorporating this infrastructure is to evaluate its contribution to overall response latency within the proposed retrieval-augmented customer support pipeline rather than to isolate the effects of hardware acceleration independently.

E. Experimental Environment

To evaluate the practical performance of the proposed architecture, two deployment configurations were considered in this study: a local GPU-based deployment and a Groq-based deployment. The proposed configuration utilized the Llama-3.3-70B-Versatile model served through Groq's cloud inference platform, while the baseline configuration employed a locally deployed language model served through Ollama and

```
System: You are an expert Customer Support Agent for
Crystallure.

Instructions:
1. Answer the user's question based ONLY on the information
provided below.
2. Read all information carefully before answering.
3. If the answer is not present in the information, strictly
reply: "Maaf, tidak ada informasi yang tersedia untuk
menjawab pertanyaan ini."
4. Do not use outside knowledge or training data.
5. Extract exact values (ml, gram, gr, g, Rp, %, etc.) when
present.
6. For product list questions, extract all unique product
names from the metadata.
7. For ingredient questions, provide complete ingredient
lists from the information.
8. Use conversation context to understand references like
"itu", "tadi", "sebelumnya" when available.

INFORMASI:
{retrieved_chunks}

PERTANYAAN: {user_query}

KONTEKS PERCAKAPAN:
{conversation_context}

INSTRUKSI:
- Cari angka/nilai yang tepat (ml, gram, gr, g, Rp, %, dll)
- Jika ada angka di awal kalimat seperti "12 gr" atau "150
ml", SEBUTTKAN angka tersebut
- Contoh: "12 gr Bedak..." berarti beratnya adalah 12 gr
- Untuk pertanyaan "apa aja produk crystallure", CARI semua
nama produk yang unik dari informasi yang diberikan dan buat
daftar lengkap
- Untuk pertanyaan "ingredientsnya" atau "kandungannya",
berikan daftar lengkap ingredients dari produk yang sedang
dibicarakan
- Jika ada konteks percakapan, gunakan untuk memahami
referensi seperti "itu", "tadi", "sebelumnya"
- Berikan jawaban yang natural dan mengalir berdasarkan
konteks percakapan
- PENTING: Untuk daftar produk, ekstrak semua nama produk
unik dari metadata "Product" yang ada di informasi
- JANGAN katakan "Informasi tidak tersedia" kecuali benar-
benar tidak ada informasi sama sekali

JAWABAN:
```

Fig. 3. The closed-context prompt structure.

executed on consumer-grade GPU hardware.

The local deployment environment consisted of an NVIDIA RTX 3060 GPU with 16 GB of system memory. Both deployment configurations utilized the same retrieval pipeline, including the Pinecone vector database, document chunking strategy, retrieval parameters, and prompt construction mechanism. This configuration was intended to ensure that the retrieval process remained consistent across experiments, allowing performance differences to be evaluated at the deployment level rather than the retrieval level.

For retrieval, the system employed a Top-K configuration of $K = 15$. Retrieved document chunks were incorporated into a closed-context prompt before response generation. All responses were generated using a temperature setting of 0 and a maximum output length of 600 tokens. These parameters were maintained throughout the evaluation process to ensure consistency in both latency and response quality measurements.

It should be noted that the comparison was designed to evaluate the practical latency characteristics of two deployment approaches within the same Retrieval-Augmented Generation workflow. The objective was not to isolate the contribution of hardware acceleration independently, but rather to assess the overall responsiveness of a locally deployed GPU-based configuration compared with the proposed Groq-based deployment in an interactive customer support scenario.

F. Evaluation Framework

The proposed system was evaluated using a domain-specific benchmark consisting of 50 question-answer pairs derived from organizational product information and customer support materials. The evaluation dataset was designed to represent common inquiry categories encountered in FMCG customer support scenarios, including product specifications, ingredients, usage instructions, and general product-related information.

System performance was assessed using two primary dimensions: response accuracy and response latency. Response accuracy was evaluated through manual assessment by comparing generated responses against the corresponding source documents contained within the knowledge base. Responses were classified as correct when the generated answer was factually consistent with the available product information and incorrect when factual inconsistencies or unsupported information were identified.

Latency evaluation focused on end-to-end response time, defined as the elapsed duration between query submission and receipt of the final generated response. Additional measurements included Time-to-First-Token (TTFT) and token generation throughput to assess inference efficiency. Latency measurements were collected under the deployment configurations described in Section III E.

The evaluation dataset was derived from a single FMCG brand and, therefore, reflects the characteristics of a specific industrial case-study setting. Consequently, the reported results should be interpreted as evidence of the practical feasibility of the proposed architecture rather than as a generalized assessment of performance across all FMCG customer support environments. Furthermore, the evaluation was conducted using only 50 test queries and relied on a single evaluator for manual response assessment. Statistical significance testing and inter-rater reliability analysis were not performed within the scope of this study. These limitations should be considered when interpreting the reported results.

Although retrieval quality plays an important role in overall RAG performance, this study did not explicitly evaluate retrieval effectiveness using metrics such as Recall@K, Precision@K, Mean Reciprocal Rank (MRR), or context coverage. Instead, the evaluation focused on end-to-end response accuracy and system latency to assess the practical feasibility of the proposed architecture in a customer support setting. Consequently, the contribution of the retrieval component was evaluated indirectly through overall response quality rather than through dedicated retrieval benchmarks.

IV. Results and Discussion

A. Response Latency Evaluation

To assess the responsiveness of the proposed architecture, a latency comparison was conducted between a local GPU-based deployment and the proposed Groq-based deployment. The evaluation focused on three metrics: Time-to-First-Token (TTFT), token generation throughput, and end-to-end response

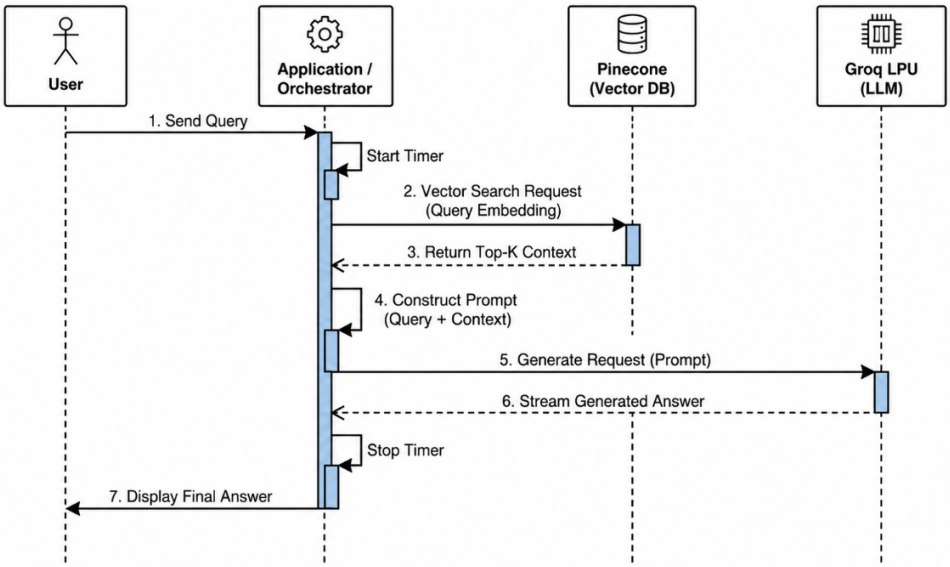


Fig. 4. Sequence diagram of the inference request.

latency. TTFT measures the time required for the first generated token to be produced, throughput measures the token generation rate, and end-to-end latency represents the total elapsed time between query submission and receipt of the final response.

Table I summarizes the latency characteristics observed for both deployment configurations. The local deployment utilized Ollama running on an NVIDIA RTX 3060 GPU, while the proposed deployment employed the Llama-3.3-70B-Versatile model served through Groq's cloud inference platform.

TABLE I. LATENCY COMPARISON BETWEEN LOCAL GPU-BASED AND GROQ-BASED DEPLOYMENTS.

Metric	Local GPU-Based	Groq-Based
Embedding & Retrieval	245 ms	238 ms
Time-to-First-Token (TTFT)	742 ms	118 ms
Token Generation Rate	42.8 tokens/sec	305.9 tokens/sec
Total Response Time	**2847 ms**	**1246 ms**

B. Response Accuracy Evaluation

The response accuracy evaluation was conducted using 50 domain-specific customer support queries derived from organizational product information. Generated responses were manually compared against the corresponding source documents contained within the knowledge base. Responses were classified as correct when the generated answer was factually consistent with the available product information and incorrect when factual inconsistencies or unsupported information were identified.

Table II summarizes the overall response accuracy obtained during the evaluation. Out of 50 evaluated queries, 36 responses were classified as correct, while 14 responses contained inaccuracies or incomplete information. This resulted in an overall response accuracy of 72.0%.

The results indicate that the proposed Retrieval-Augmented Generation architecture was able to provide factually consistent responses for the majority of evaluated customer support queries. The integration of retrieval-based grounding enabled the language model to access product-specific information that was not explicitly contained within its parametric knowledge.

TABLE II. RESPONSE ACCURACY RATE EVALUATION RESULTS

Evaluation Outcome	Count
Correct Responses	36
Incorrect Responses	14
Total Queries	50
Accuracy	**72.0%**

Analysis of incorrect responses revealed that most errors originated from incomplete retrieval results, ambiguity within product descriptions, or situations where relevant information was available in the retrieved context but was not fully incorporated into the generated response. These findings suggest that retrieval quality and context utilization remain important factors influencing overall system performance.

C. Discussion

The experimental results demonstrate that retrieval augmentation can improve the ability of large language models to answer domain-specific customer support queries. The proposed system achieved an overall response accuracy of 72.0% across the evaluated FMCG customer support dataset, indicating that retrieval-based grounding can assist the language model in generating responses that are consistent with organizational product information.

The results suggest that integrating an external knowledge repository enables the language model to access product-specific information that is not explicitly stored within its parametric knowledge. By incorporating retrieved document chunks into the generation process, the system can provide responses grounded in available product documentation rather than relying solely on pre-trained model knowledge.

The latency evaluation further indicates that the Groq-based deployment provides faster response times than the local GPU-based deployment. As shown in Table I, the proposed deployment reduced Time-to-First-Token (TTFT) from 742 ms to 118 ms and decreased end-to-end response latency from 2847 ms to 1246 ms. In addition, token generation throughput increased from 42.8 tokens per second to 305.9 tokens per second. These findings suggest that low-latency inference infrastructure can improve the responsiveness of retrieval-augmented customer support systems.

Despite the observed improvements, the retrieval component remains an important source of performance limitations. The current implementation employs a lightweight hash-based vector representation that primarily captures lexical similarity rather than deep semantic relationships between terms. Consequently, retrieval effectiveness may decrease when semantically similar queries are expressed using substantially different vocabulary. This limitation may partially explain some of the incorrect responses observed during evaluation.

Analysis of incorrect responses revealed that most errors originated from incomplete retrieval results, ambiguity within product descriptions, or situations where relevant information was available in the retrieved context but was not fully incorporated into the generated response. These findings indicate that retrieval quality and context utilization remain important factors influencing overall system performance.

From an operational perspective, the proposed architecture enables organizational knowledge to be updated independently from the language model itself. New product information can be incorporated by updating the knowledge base and vector index without requiring model retraining. This characteristic may be beneficial in FMCG environments where product information and operational procedures change frequently and require timely updates across customer support channels.

D. Limitations and Future Work

Several limitations should be considered when interpreting the findings of this study. First, the evaluation was conducted using a dataset consisting of 50 question-answer pairs derived from product information belonging to a single FMCG brand. Although the selected dataset represents a realistic industrial use case, the reported results may not generalize to larger knowledge repositories, multiple brands, or different organizational environments. Furthermore, response quality was assessed through manual evaluation and did not involve multiple independent annotators. Consequently, inter-rater reliability was not measured, and the evaluation dataset of only 50 query-response pairs was insufficient to support meaningful statistical significance testing. Future studies should expand the evaluation benchmark and incorporate multiple independent annotators to improve evaluation reliability and generalizability.

A second limitation relates to the retrieval component. The current implementation employs a lightweight hash-based vector representation that primarily captures lexical similarity rather than deep semantic relationships between terms. While computationally efficient, this approach may reduce retrieval effectiveness when semantically similar queries are expressed using substantially different vocabulary. In addition, the study focused on end-to-end system performance and did not evaluate retrieval effectiveness using dedicated retrieval metrics such as Recall@K, Precision@K, Mean Reciprocal Rank (MRR), or context coverage. Consequently, the contribution of the retrieval component was assessed indirectly through overall response quality rather than through dedicated retrieval benchmarks.

Controlled ablation experiments were also not conducted within the scope of this study due to the limited experimental setup and available evaluation data. Consequently, the individual contributions of retrieval configuration, prompt design, serving infrastructure, and inference hardware could not be isolated quantitatively. Future work should perform systematic ablation studies to better understand the relative impact of each component on overall system performance.

Finally, the study did not investigate scalability under concurrent user workloads, larger knowledge base sizes, or alternative deployment frameworks such as vLLM and TensorRT-LLM. Future research should address these aspects while also exploring transformer-based embedding models, hybrid retrieval strategies, and reranking techniques to improve retrieval quality and response accuracy. Such investigations would provide a more comprehensive understanding of deployment-oriented Retrieval-Augmented Generation systems in industrial customer support environments.

V. Conclusion

This study presented the design and evaluation of a Retrieval-Augmented Generation (RAG) architecture for domain-specific customer support within the Fast-Moving Consumer Goods (FMCG) sector. The proposed system integrates document preprocessing, vector-based retrieval, closed-context prompting, and large language model inference to generate responses grounded in organizational product information. In addition, the study evaluated two deployment configurations, namely a local GPU-based deployment and a Groq-based deployment, to investigate their practical latency characteristics within the same retrieval-augmented workflow.

The experimental evaluation demonstrated that the proposed system achieved an overall response accuracy of 72.0% across 50 domain-specific customer support queries. The results indicate that retrieval-based grounding can assist large language models in providing responses that are more consistent with organizational knowledge sources. In terms of system responsiveness, the Groq-based deployment reduced end-to-end response latency from 2847 ms to 1246 ms, improved Time-to-First-Token (TTFT) from 742 ms to 118 ms, and increased token generation throughput from 42.8 to 305.9 tokens per second. These findings suggest that low-latency inference infrastructure can support responsive customer support interactions while maintaining retrieval-grounded response generation.

The findings should be interpreted within the scope of an industrial case-study setting involving product information from a single FMCG brand. Furthermore, the evaluation dataset consisted of a limited number of manually assessed queries and did not include dedicated retrieval effectiveness metrics, multiple independent annotators, or large-scale scalability experiments. Consequently, the reported results should

be viewed as evidence of practical feasibility rather than a generalized assessment of Retrieval-Augmented Generation performance across all enterprise environments.

Future work should investigate transformer-based embedding models, retrieval effectiveness metrics such as Recall@K, Precision@K, and Mean Reciprocal Rank (MRR), as well as hybrid retrieval and reranking approaches to improve retrieval quality. Additional studies should also evaluate scalability under larger knowledge bases, concurrent user workloads, and alternative deployment frameworks to provide a more comprehensive understanding of deployment-oriented Retrieval-Augmented Generation systems in industrial customer support applications.

ACKNOWLEDGMENT

We express our gratitude to the PT Paragon Technology and Innovation for facilitating this case study, particularly Mrs. Nurul Rizka Hadiningsih for their industrial mentorship. We also appreciate the valuable input provided by the Customer Care team leaders and agents during the system's development and testing.

REFERENCES

[1] S. R. Peddinti, S. R. Katragadda, B. K. Pandey, and A. Tanikonda, "Utilizing large language models for advanced service management: potential applications and operational challenges," *Journal of Science & Technology*, vol. 4, no. 2, 2023.

[2] S. Veturi, S. Vaichal, R. L. Jagadheesh, N. I. Tripto, and N. Yan, "Rag based question-answering for contextual response prediction system," *arXiv preprint arXiv:2409.03708*, 2024.

[3] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[4] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

[5] A. G. Larsen, M. B. Skjuve, A. Følstad, and N. Van As, "Llm hallucinations in conversational ai for customer service: Framework and end-user perceptions," *International Journal of Human–Computer Interaction*, pp. 1–22, 2025.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[7] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[8] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.

[9] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," in *International conference on machine learning*. PMLR, 2023, pp. 15 696–15 707.

[10] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th symposium on operating systems principles*, 2023, pp. 611–626.

Adaptive Hybrid Intrusion Detection for Realistic Zero-Day Attacks in Cloud and Edge Environments

Nithin U, Ganeshayya Shidaganti*, Sangeetha V, Vishwachetan D
Department of Computer Science and Engineering,
Ramaiah Institute of Technology, Bangalore, Karnataka, India

***Abstract*—New and unknown attack patterns are creating more cybersecurity issues for cloud environments. Intrusion detection systems (IDS) are usually capable of high-performance in closed-world scenarios and are less effective in realistic zero-day scenarios. This work presents an adaptive hybrid intrusion detection mechanism for cloud environments based on the combination of Random Forest, XGBoost and AutoEncoder models, under a single decision-making framework. A fully reproducible pipeline was built directly from the raw data of CICIDS2017, using consistent preprocessing and feature engineering. In contrast with traditional IDS researches that are evaluated with standard train-test evaluation process, realistic zero-day assessment was carried out by conducting attack-family holdout experiments, including WebAttack, DDoS, and Infiltration scenarios. The results showed that both attack families and adaptive decision strategies were different, with WebAttack using XGBoost-dominant behavior and Infiltration using anomaly-focused behavior for an 85.30% and 88.89% recall, respectively. The proposed framework was also implemented on edge hardware (Raspberry Pi) and has been shown to have low-inference latency and only moderate resource usage. The results reveal that adaptive hybrid decision behavior leads to a more robust decision behavior under realistic zero-day conditions, while having a feasible deployment in cloud-edge environments.**

Keywords*—*Intrusion detection system (IDS); zero-day detection; adaptive hybrid IDS; attack-family holdout; cloud security; edge deployment; autoencoder; CICIDS2017

I. Introduction

Cloud computing has become a fundamental component of modern computing infrastructure because of its scalability, flexibility, and efficient resource utilization. As organizations increasingly rely on cloud platforms for large-scale services and distributed applications, cybersecurity challenges have also grown significantly. Evolving attack behaviors and large-scale network traffic generation make securing cloud environments increasingly difficult. Recent studies have emphasized the importance of intelligent intrusion detection systems capable of adapting to dynamic cloud environments [1], [2].

Intrusion detection systems (IDS) are commonly used to monitor network traffic and identify malicious activity. Traditional IDS approaches primarily rely on signature-based techniques or supervised machine learning models trained using historical attack data. Although such approaches achieve strong performance on known attack categories [3], their effectiveness decreases when exposed to previously unseen threats. Zero-day attacks remain particularly challenging because attack signatures are unavailable during training, limiting the generalization capability of conventional models [4], [5].

*Corresponding author

A major limitation in existing intrusion detection studies is the reliance on conventional train-test evaluation protocols where attack distributions remain similar across training and testing stages. Such closed-world assumptions frequently yield high classification performance, but are not representative of reality deployments in scenarios with new attack families not encountered before. Thus, models with good accuracy when evaluated in standard settings can suffer a significant drop in accuracy when evaluated in atypical settings of realistic zero day conditions.

An adaptive hybrid IDS framework for realistic zero-day attacks in the cloud can be used to overcome these challenges. To achieve this, a new method is proposed that integrates three machine learning algorithms (Random Forest, XGBoost and Autoencoder) into a single adaptive decision-making process. Contrary to previous researches that have been done in a closed world setting, here a new adaptive hybrid decision-making process is introduced which has been verified under realistic attack-family based holdouts and edge deployment. Zero-day evaluation has been completed for the WebAttack, DDoS and Infiltration holds and Raspberry Pi deployment for inference viability.

The major contributions of this work are summarized as follows:

- A reproducible intrusion detection pipeline was reconstructed directly from the raw CICIDS2017 dataset using consistent preprocessing, feature engineering, and model training procedures.
- A realistic zero-day evaluation strategy based on attack-family holdout experiments was developed using WebAttack, DDoS, and Infiltration scenarios.
- An adaptive hybrid intrusion detection mechanism combining Random Forest, XGBoost, and Autoencoder models was developed to improve robustness across diverse attack behaviors.
- Raspberry Pi edge deployment and dashboard-based visualization were implemented to validate practical deployment feasibility.

II. Related Work

A. Deep Learning and Hybrid Intrusion Detection

Recent intrusion detection studies have increasingly adopted deep learning techniques to improve the detection of complex and previously unseen attack patterns. Sotani et al. [6] demonstrated that deep learning architectures can improve

network traffic analysis compared with conventional methods. To improve robustness further, Aljawarneh et al. [7] proposed a hybrid intrusion detection architecture combining multiple learning paradigms. Cloud-specific intrusion detection studies have also emphasized scalable and adaptive monitoring in dynamic environments. AI-based cloud security frameworks [8] and hybrid machine learning-deep learning approaches [9] have reported improved detection capability under cloud-oriented settings. However, these methods mainly focus on predefined attack categories and continue to rely on labeled attack information.

B. Adaptive and Zero-Day Detection Approaches

To address limited data availability in unseen attack scenarios, few-shot learning has been explored for intrusion detection. Yan et al. [10] showed that few-shot frameworks can learn new attack categories using limited labeled samples. Zero-shot learning has also been investigated for unseen attack identification by transferring knowledge from known attack classes to unknown classes [11]. Recent AI-driven hybrid methods have attempted to improve zero-day exploit detection by combining multiple classifiers and deep learning models [12]. Time-series anomaly detection approaches have further improved the ability to capture dynamic cloud traffic behavior [13], while ensemble-based hybrid techniques have demonstrated improved cyberattack detection performance [14]. Meta-learning approaches have also been applied to intrusion detection in next-generation network environments [15].

C. Research Gap

Earlier deep learning-based zero-day detection studies reported improved representation learning capability but remained largely dependent on supervised or controlled evaluation settings [16]. Simulation-based zero-day studies showed that detection performance decreases when models encounter unseen attack distributions [17]. Recent survey work has also highlighted persistent challenges such as false positives, adaptability limitations, and lack of practical deployment validation [18]. Cloud-focused zero-day studies have proposed deep learning and machine learning frameworks for emerging threats [19], [20]. However, most existing studies continue to rely on conventional dataset partitioning or closed-world assumptions. Realistic attack-family holdout evaluation, adaptive hybrid decision behavior, and practical edge deployment validation remain comparatively underexplored. Unlike existing approaches, this work combines adaptive hybrid decision logic with realistic attack-family holdout experiments and Raspberry Pi edge deployment validation.

III. Proposed Methodology

A. Overall System Architecture

The proposed system is designed as an adaptive hybrid intrusion detection framework capable of detecting both known and previously unseen attacks in cloud environments. The architecture combines supervised classification and anomaly-based learning within a unified detection pipeline, as illustrated in Fig. 1. Unlike conventional IDS frameworks that depend entirely on labeled attack information, the proposed approach integrates multiple detection paradigms to improve robustness under realistic zero-day conditions.

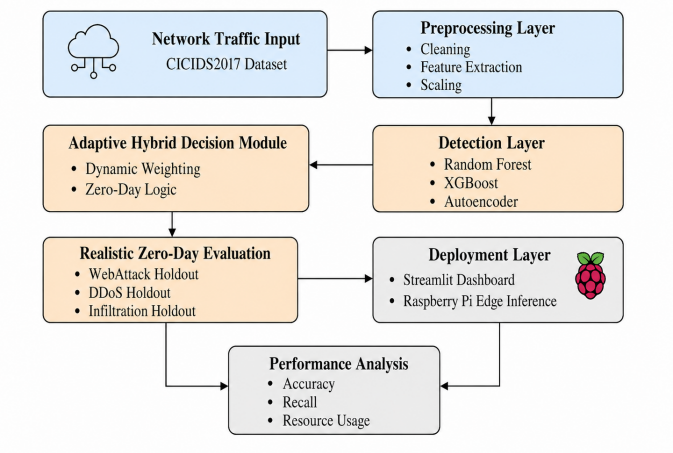


Fig. 1. Overall architecture of the proposed adaptive hybrid IDS framework for realistic zero-day detection and edge deployment.

The entire process starts with the processing of network traffic data and the extraction of the features. The preprocessed network flow features are then processed by both supervised and anomaly-based detectors at the same time. The Random Forest and the XGBoost model learn the discriminative features from the training samples that represent attack traffic, whereas the Autoencoder learns the normal traffic by training only on benign traffic data. The outputs from the two models are then fed into an adaptive decision making unit which is responsible for making final decisions about the intrusions.

B. Dataset Preparation and Preprocessing

The proposed framework utilized the CICIDS2017 dataset as the primary source of network traffic data. To maintain reproducibility and eliminate inconsistencies associated with previously processed artifacts, only raw CSV files were used throughout experimentation. Initially, raw traffic files were merged into a unified dataset and standardized through column cleaning and duplicate removal. Missing and infinite values were identified and handled to ensure numerical stability during model training.

Feature engineering was then performed to retain only numerical attributes suitable for model learning. Labels were converted into binary form, where benign traffic was assigned class 0 and malicious traffic was assigned class 1. A final set of 78 traffic features was selected and preserved consistently across training and inference stages. The dataset was divided into training, validation, and testing subsets using a 70:15:15 split strategy, and StandardScaler was applied using training data only to prevent information leakage.

C. Hybrid Intrusion Detection Components

1) Random Forest model: Random Forest was employed as a supervised learning model for known attack classification. Random Forest constructs multiple decision trees and combines predictions through ensemble voting. Due to its

robustness against overfitting and ability to process high-dimensional traffic features efficiently, it serves as an effective baseline classifier for intrusion detection. The Random Forest component primarily contributes toward identifying attack behaviors already represented during training.

2) XGBoost model: XGBoost was incorporated to improve discriminative capability through gradient boosting. Unlike Random Forest, XGBoost progressively corrects prediction errors during training and learns complex relationships within network traffic patterns. Experimental observations showed that XGBoost demonstrated stronger performance under specific zero-day scenarios such as WebAttack experiments. Therefore, it played an important role within the adaptive decision framework.

3) Autoencoder-based anomaly detection: An Autoencoder was employed to model normal network behavior using benign traffic samples only. The Autoencoder architecture consisted of a symmetric encoder-decoder structure designed to learn compact representations of benign network traffic behavior.

The encoder progressively reduced feature dimensions from $78 \rightarrow 64 \rightarrow 32 \rightarrow 16$, while the decoder reconstructed traffic representations using $16 \rightarrow 32 \rightarrow 64 \rightarrow 78$. This structure enabled learning of compressed latent representations while preserving important traffic characteristics required for anomaly detection.

Let the input feature vector be represented as:

$$x \in R^d \tag{1}$$

The encoder transforms input data into a latent representation:

$$z = f(x) \tag{2}$$

The decoder reconstructs the original input:

$$\hat{x} = g(z) \tag{3}$$

The reconstruction loss is calculated using mean squared error:

$$L(x, \hat{x}) = \frac{1}{d} \sum_{i=1}^{d} (x_i - \hat{x}_i)^2 \tag{4}$$

Higher reconstruction error values indicate stronger deviation from learned normal behavior and are therefore treated as anomaly indicators. Unlike supervised classifiers, the Autoencoder provides sensitivity toward previously unseen attack behavior.

D. Adaptive Hybrid Decision Logic

The final intrusion decision is generated through adaptive hybrid fusion of Random Forest, XGBoost, and Autoencoder outputs. The hybrid score is computed as:

$$H(x) = w_1 RF(x) + w_2 XGB(x) + w_3 AE(x) \tag{5}$$

where,

$$w_1 + w_2 + w_3 = 1 \tag{6}$$

The final intrusion decision was determined using a threshold-based decision rule:

$$Decision(x) = \begin{cases} Attack, & \text{if } H(x) > T \\ Benign, & \text{otherwise} \end{cases} \tag{7}$$

where, T represents the decision threshold used to distinguish benign and malicious network behavior. The hybrid decision formulation provides a unified representation for combining the outputs of Random Forest, XGBoost, and Autoencoder models. In this study, attack-family holdout experiments were used to analyze how individual detection components responded to different unseen attack characteristics. Experimental observations showed that WebAttack scenarios exhibited stronger responses from the XGBoost component, Infiltration attacks were more effectively identified through Autoencoder-based anomaly behavior, and DDoS experiments demonstrated relatively balanced contributions from the detection components.

Therefore, the adaptive behavior discussed in this work represents an experimental analysis of the contribution of supervised and anomaly-based models under different unseen attack scenarios rather than a deployment-time mechanism that automatically identifies attack families and dynamically selects optimal weights. These observations highlight the limitations of using a single fixed detection emphasis across diverse zero-day attack behaviors and motivate further investigation toward fully automated adaptive decision mechanisms.

1) Hyperparameter configuration: To ensure reproducibility and stable model behavior, experimentally selected hyperparameters were used during training and evaluation. The configuration settings used for Random Forest, XGBoost, and Autoencoder training are summarized in Table I.

TABLE I. HYPERPARAMETER CONFIGURATION USED DURING MODEL TRAINING.

Model	Configuration
Random Forest	n_estimators = 100, max_depth = 20
XGBoost	n_estimators = 100, learning rate = 0.1, max depth = 6
Autoencoder	Encoder: 78-64-32-16, Batch Size = 256, Epochs = 20, Optimizer = Adam

E. Realistic Zero-Day Evaluation Design

Conventional intrusion detection evaluations commonly rely on random train-test splitting, where attack distributions remain similar across training and testing stages. Such protocols may produce inflated performance estimates and fail to represent practical deployment conditions. To create realistic zero-day scenarios, complete attack families were removed

from training and introduced only during testing. Multiple experiments were performed using attack-family holdout methodology.

The following experiments were conducted:

- Experiment 1: WebAttack holdout.
- Experiment 2: DDoS holdout.
- Experiment 3: Infiltration holdout.

This strategy ensured that selected attack behaviors remained entirely unseen during training.

IV. Experimental Setup and Evaluation Protocol

A. Dataset Description

The proposed intrusion detection framework was evaluated using the CICIDS2017 dataset, which is widely used for benchmarking network intrusion detection systems under modern traffic conditions. The dataset contains both benign and malicious network traffic generated under realistic communication scenarios. For this work, only the raw CICIDS2017 CSV files were used to ensure reproducibility and eliminate inconsistencies introduced by previously processed datasets.

After merging and preprocessing, approximately 2.57 million traffic samples were retained. Network flows were represented using 78 numerical features extracted from traffic statistics and protocol-level characteristics. The dataset characteristics and attack-family distributions used in this work are summarized in Table II.

TABLE II. Summary of CICIDS2017 Dataset Characteristics and Preprocessing.

Parameter	Value
Dataset	CICIDS2017
Raw CSV files	8
Samples after preprocessing	∼2.57 million
Features used	78
Labels	Benign = 0, Attack = 1
Train : Test : Val	70 : 15 : 15
BENIGN samples	2,146,899
DDoS samples	128,014
WebAttack samples	2,143
Infiltration samples	36

B. Experimental Environment

Experiments were conducted using a combination of local and cloud-based environments to support large-scale model training and deployment analysis. Model development and experimentation were performed using Python with machine learning and deep learning libraries, including Scikit-learn, XGBoost, PyTorch, NumPy, and Pandas. Initial experimentation and training activities were conducted using Google Colab and later migrated to a local development environment to improve reproducibility and avoid computational limitations.

The local system configuration consisted of:

- AMD Ryzen 5 5600H processor
- NVIDIA GTX 1650 GPU
- 16 GB RAM
- VS Code development environment

To evaluate practical deployment feasibility, inference experiments were additionally conducted using Raspberry Pi edge hardware. The edge system was used exclusively for model loading, adaptive inference, resource monitoring, and latency analysis.

C. Evaluation Protocol

The proposed system was evaluated under multiple experimental settings to assess behavior under both conventional and realistic intrusion scenarios. The first evaluation stage consisted of baseline closed-world testing using standard train-test distributions. The second evaluation stage focused on realistic zero-day assessment using attack-family holdout experiments. Unlike random train-test splitting approaches, complete attack categories were removed during training and introduced only during testing. Training datasets contained benign traffic and all remaining attack categories except the selected holdout family.

D. Ablation Study Design

To analyze the contribution of individual components within the proposed framework, an ablation study was conducted. The following configurations were evaluated independently: Random Forest only, XGBoost only, Autoencoder only, and the Hybrid framework. This analysis enabled identification of strengths and weaknesses associated with supervised, anomaly-based, and combined detection approaches. The performance comparison of individual models and the hybrid framework is presented in Table III.

TABLE III. Ablation Performance Comparison of RF, XGBoost, Autoencoder, and Hybrid IDS.

Model	Accuracy	Precision	Recall	F1
RF	0.9985	0.9966	0.9947	0.9957
XGBoost	0.9992	0.9969	0.9986	0.9977
Autoencoder	0.8890	0.7065	0.5871	0.6413
Hybrid	0.9990	0.9974	0.9967	0.9970

The Autoencoder produced lower closed-world accuracy because it was trained using benign traffic only and optimized for anomaly detection. The hybrid framework improved robustness by combining supervised learning capability with anomaly-sensitive behavior.

E. Performance Metrics

Performance evaluation was conducted using commonly accepted intrusion detection metrics. Accuracy was used to measure overall classification correctness. Precision quantified the proportion of predicted attacks that were correctly identified, while recall measured the capability of detecting malicious samples. F1-score provided a balanced measure combining precision and recall. In addition, Receiver Operating Characteristic (ROC) analysis and Area Under Curve (AUC) values were used. For edge deployment experiments, additional deployment metrics were measured: prediction latency, CPU utilization, and memory consumption.

V. Results and Discussion

A. Baseline Performance Evaluation

The initial evaluation was performed under conventional closed-world settings where attack categories present during testing were also represented during training. The supervised Random Forest and XGBoost models achieved very high performance across standard evaluation metrics. The hybrid framework combined outputs from all components and maintained strong overall performance. The observed results indicate that supervised learning approaches perform extremely well when attack distributions remain similar across training and testing datasets. However, such results should be interpreted carefully because high closed-world accuracy does not necessarily indicate effective generalization capability under realistic deployment conditions.

B. Realistic Zero-Day Evaluation Results

To assess realistic generalization capability, multiple attack-family holdout experiments were performed (WebAttack, DDoS, and Infiltration). Experimental observations revealed substantial differences in model behavior across attack categories. Detection capability varied significantly depending on traffic characteristics and attack structure. The performance comparison across realistic zero-day experiments is summarized in Table IV.

TABLE IV. Performance Comparison Under Realistic Zero-Day Attack-Family Holdout Experiments.

Experiment	Accuracy	Recall	Detected	Missed	Adaptive Strategy
WebAttack	93.77	85.30	1828	315	XGB-heavy
DDoS	82.13	64.37	1931	1069	Balanced
Infiltration	94.33	88.89	32	4	AE-dominant

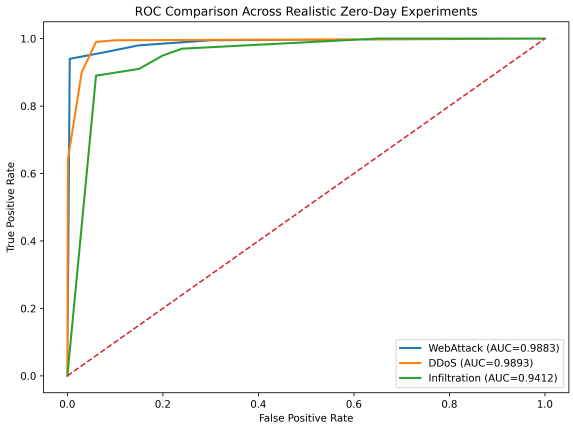


Fig. 2. ROC comparison across realistic zero-day attack-family holdout experiments.

The ROC analysis (Fig. 2) further validates discrimination capability across the realistic attack-family scenarios. WebAttack and DDoS experiments achieved strong AUC behavior, while Infiltration demonstrated comparatively lower separation characteristics because of extremely limited sample availability.

C. Adaptive Hybrid Analysis

Experimental evaluation under attack-family holdout scenarios revealed that the individual detection components exhibited different responses to unseen attack characteristics. The relative contribution of Random Forest, XGBoost, and Autoencoder components across the evaluated scenarios is illustrated in Fig. 3. These observations indicate that relying on a single detection emphasis may not provide consistent behavior across diverse zero-day attack patterns.

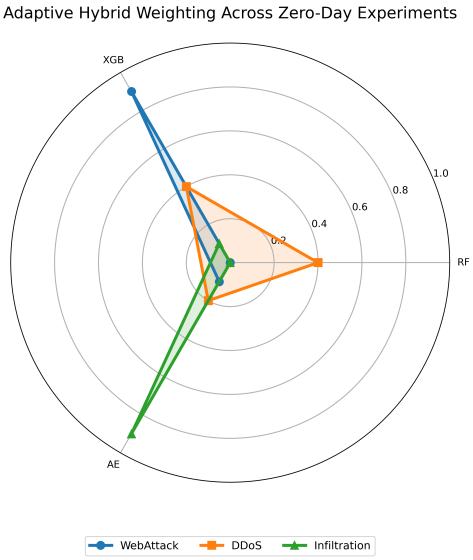


Fig. 3. Relative contribution of Random Forest, XGBoost, and Autoencoder components across realistic zero-day attack-family experiments.

For the WebAttack holdout experiment, the XGBoost component demonstrated stronger generalization behavior and contributed significantly to the final hybrid decision, resulting in the successful detection of 1828 out of 2143 attack samples with a recall of 85.30%.

In the DDoS holdout experiment, the detection components exhibited relatively balanced behavior, indicating that both supervised and anomaly-based characteristics contributed to the overall decision process. However, the achieved recall of 64.37% was noticeably lower than the performance observed for WebAttack and Infiltration scenarios. Since the DDoS attack family was entirely excluded during training, this result highlights the difficulty of generalizing intrusion detection models to certain unseen attack behaviors under realistic attack-family holdout conditions. These observations indicate that zero-day detection performance may vary substantially across different attack families and emphasize the importance of evaluating intrusion detection systems beyond conventional closed-world testing.

The Infiltration scenario represented a particularly challenging case where the supervised models showed limited detection capability, whereas the Autoencoder exhibited stronger anomaly-sensitive behavior. The hybrid analysis identified this difference in model response, resulting in the detection of 32 out of 36 Infiltration samples with a recall of 88.89%. However, this result should be interpreted carefully because

the limited number of available Infiltration samples in CI-CIDS2017 restricts broad generalization of this observation.

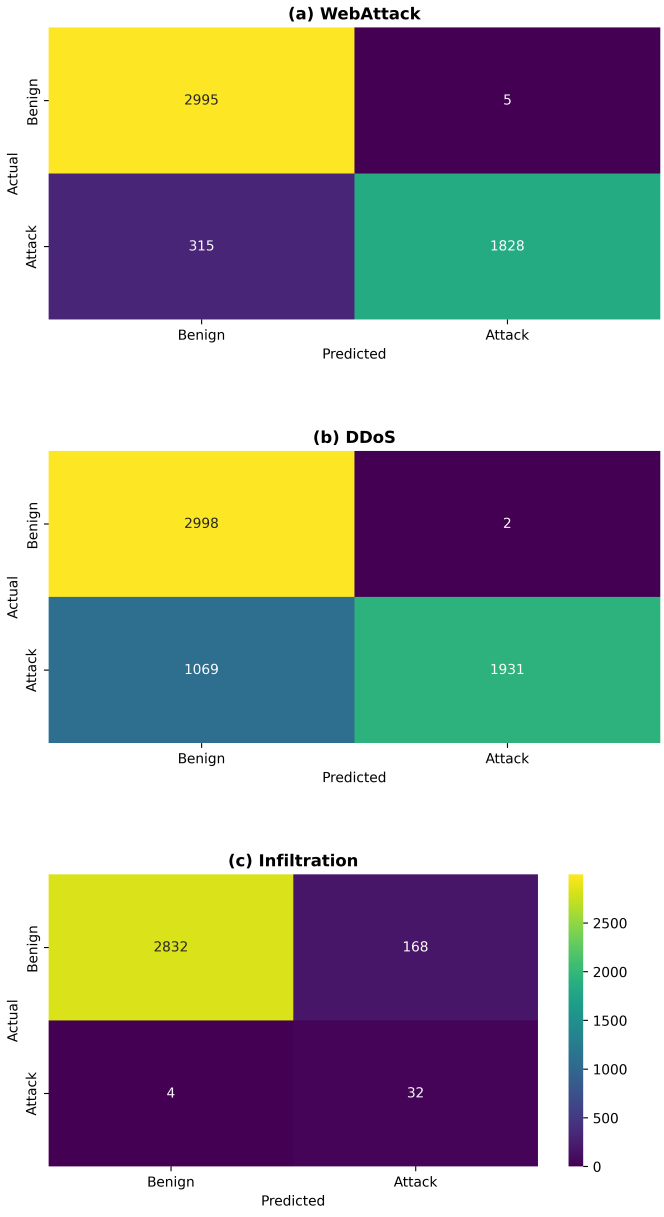


Fig. 4. Confusion matrices for realistic zero-day attack-family experiments.

The classification behavior across the three attack-family holdout experiments is further illustrated through the confusion matrices shown in Fig. 4. WebAttack and DDoS experiments achieved comparatively stronger detection performance, while the Infiltration experiment remained challenging because of the limited sample availability and its distinct attack behavior. Fig. 5 presents the score distributions obtained from the hybrid decision analysis across different attack-family holdout scenarios. WebAttack and DDoS produced score distributions concentrated around moderate decision regions, whereas the Infiltration scenario showed a stronger anomaly-oriented response with scores distributed closer to the higher decision region.

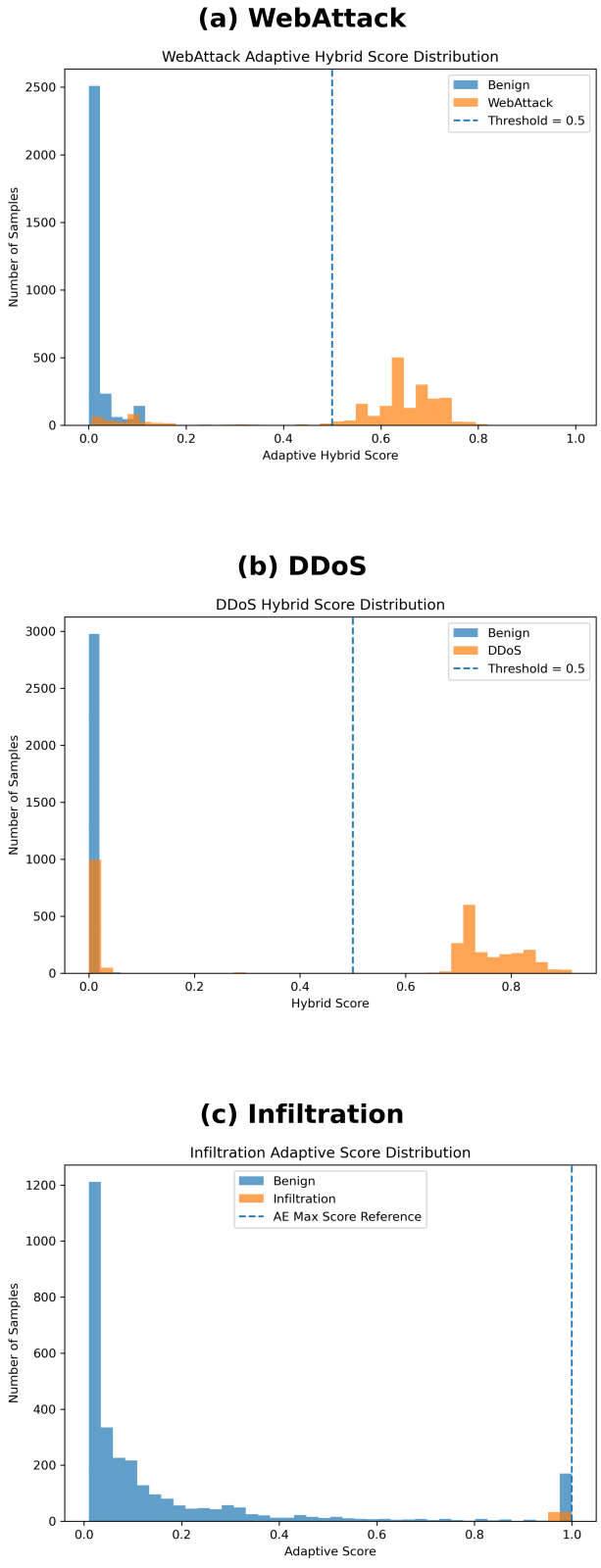


Fig. 5. Adaptive score distributions across realistic zero-day attack-family experiments.

D. Discussion of Key Findings

Several important findings emerged from experimental evaluation. The first is that closed world accuracy values close to 100% can give false readings for an intrusion detection system. While supervised models were highly successful under standard evaluation conditions, their performance significantly dropped when realistic unseen attacks were added.

Secondly, the attack-family holdout experiments demonstrated that different unseen attack scenarios produced different responses from the individual detection components. XGBoost showed stronger behavior in the WebAttack holdout experiment, whereas the Infiltration holdout experiment indicated greater sensitivity from the Autoencoder-based anomaly detection component. However, the Infiltration observations should be interpreted within the limitations of the CICIDS2017 dataset, where only 36 samples were available for evaluation.

Thirdly, observed during the attack-family holdout experiments are the different reactions to unknown attack scenarios by the individual detection elements. This observation suggests that a unique detection emphasis may not provide stable behavior across different kinds of zero-day attacks. This was proven through experimental results, which indicates that realistic attack-family holdout evaluation provides more information than traditional closed-world evaluation, the approach often utilized in intrusion detection research studies. The hybrid framework, described above, illustrates how supervised and anomaly-based detection components can exhibit complementary behavior toward unknown attack situations, advancing the state of current research in order to improve realistic zero-day intrusion detection and hybrid IDS evaluation [5], [7], [12].

E. Edge Deployment Results

To examine practical feasibility, the adaptive hybrid intrusion detection system was deployed on Raspberry Pi edge hardware. Model training remained external to the edge device, while inference and adaptive decision logic were executed directly within the Raspberry Pi environment. Experimental results demonstrated successful deployment with consistent prediction behavior under edge execution conditions. Resource monitoring showed low CPU usage, moderate memory consumption, and low prediction latency across experiments (see Table V).

TABLE V. RASPBERRY PI DEPLOYMENT PERFORMANCE AND RESOURCE UTILIZATION FOR REALISTIC ZERO-DAY EXPERIMENTS.

Experiment	Acc. (%)	Recall (%)	Latency (ms)	CPU (%)	RAM (MB)
WebAttack	94.42	86.93	0.0127	1.0	151.55
DDoS	82.13	64.37	0.0096	0.0	154.00
Infiltration	94.33	88.89	0.0122	0.0	150.98

With approximately 151 MB memory use, WebAttack deployment proved to be 94.42% accurate and had a recall of 86.93%. Therefore, it was clear that cloud-based trained adaptive intrusion detection systems can be deployed efficiently on edge devices that have lower computational power and can still maintain the capability of making inferences.

VI. CONCLUSION AND FUTURE WORK

In this work, an adaptive hybrid intrusion detection framework for realistic zero-day detection in cloud environment was presented. This work emphasized on reproducibility, realistic attack-family holdout experiments, and deployment feasibility, as opposed to traditional intrusion detection studies that come with closed-world evaluation settings. To prevent problems with preprocessing mismatch and feature inconsistency, the whole pipeline was rebuilt from raw CICIDS2017 data with a consistent pre-processing method.

The proposed solution integrated the Random Forest, XGBoost and Autoencoder models in a single adaptive decision mechanism. Experimental testing showed that closed world accuracy alone is not enough to evaluate intrusion detection capability in real world applications.Attack-family holdout experiments involving WebAttack, DDoS, and Infiltration demonstrated that different unseen attack scenarios produced varying responses from the individual detection components. The experimental observations highlight the importance of combining supervised and anomaly-based detection perspectives when evaluating realistic zero-day conditions. The Infiltration results, although showing promising anomaly-sensitive detection behavior within CICIDS2017, should be interpreted with consideration of the limited number of available evaluation samples. In addition, successful deployment on Raspberry Pi edge hardware validated the feasibility of executing adaptive intrusion detection logic on resource-constrained devices.

Future work may extend the proposed framework using live network traffic evaluation, federated learning architectures, and larger IoT-based environments. Additional investigation into dynamic threshold optimization and continuously evolving attack distributions may further improve practical zero-day detection capability.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the Department of Computer Science and Engineering, Ramaiah Institute of Technology, for providing the facilities, technical resources, and academic environment required for this research work. The authors also acknowledge the valuable guidance and continuous support provided during the development, evaluation, and deployment phases of the proposed adaptive intrusion detection framework. The authors further express their appreciation to the MSRIT Alumni Association (MSRITAA) for its support and encouragement.

REFERENCES

[1] M. Alazab, S. Venkatraman, and P. Watters, "An intrusion detection model to detect zero-day attacks in unseen data using machine learning," in *Proceedings of the IEEE International Conference on Advanced Information Networking and Applications (AINA)*, 2023, pp. 1–8.

[2] B. I. Hairab, H. K. Aslan, M. S. Elsayed, A. D. Jurcut, and M. A. Azer, "Anomaly detection of zero-day attacks based on CNN and regularization techniques," *IEEE Access*, vol. 11, pp. 1–15, 2023.

[3] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Cloud network anomaly detection using machine and deep learning," *Cluster Computing*, vol. 25, no. 4, pp. 2541–2556, 2022.

[4] A. Khraisat, I. Gondal, and P. Vamplew, "Cloud network traffic classification and intrusion detection system," in *Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2021, pp. 1–7.

[5] S. Moustafa and J. Slay, "Detecting zero-day intrusion attacks using semi-supervised machine learning," *Journal of Information Security and Applications*, vol. 67, pp. 1–12, 2022.

[6] A. Sotani, R. Kanai, and S. Ozawa, "DL-based intrusion detection system on zero-day attacks," in *Proceedings of the IEEE International Conference on Cyber Security and Cloud Computing*, 2023, pp. 1–6.

[7] S. Aljawarneh, M. Aldwairi, and M. Yassein, "Empowering intrusion detection systems: A synergistic hybrid approach," *Journal of Network and Computer Applications*, vol. 212, pp. 103–118, 2023.

[8] R. Sharma, A. Patel, and N. Gupta, "Enhancing cloud security: Implementing AI-based intrusion detection," in *Proceedings of the IEEE International Conference on Cloud Computing and Intelligence Systems*, 2022, pp. 1–6.

[9] A. A. Aburomman and M. B. I. Reaz, "Enhancing intrusion detection: A hybrid machine and deep learning approach," *IEEE Systems Journal*, vol. 17, no. 2, pp. 1–12, 2023.

[10] Y. Yan, Y. Yang, F. Shen, M. Gao, and Y. Gu, "Few-shot learning-based method for industrial internet intrusion detection," *IEEE Internet of Things Journal*, vol. 11, no. 3, pp. 1–12, 2024.

[11] M. Ring, D. Landes, A. Hotho, and A. Nissen, "From zero-shot machine learning to zero-day attack detection," *Computers & Security*, vol. 122, pp. 102–116, 2023.

[12] A. Alzahrani, M. Alenezi, and S. Almotairi, "Hybrid AI-driven techniques for enhancing zero-day exploit detection," *Future Generation Computer Systems*, vol. 148, pp. 1–14, 2024.

[13] L. Manjula and E. Saravanakumar, "Intrusion detection in cloud computing based on time-series anomalies," in *Proceedings of the IEEE International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2023, pp. 1–8.

[14] A. Verma and S. R. Choudhary, "Machine learning-based hybrid technique to enhance cyber-attack detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 1–13, 2023.

[15] Y. Li, Z. Wang, J. Liu, and X. Chen, "Meta-learning-based few-shot intrusion detection for 5G-enabled networks," *IEEE Transactions on Network and Service Management*, vol. 20, no. 4, pp. 1–12, 2023.

[16] S. Kumar, R. Ranjan, and A. K. Singh, "Utilising deep learning techniques for effective zero-day attack detection," *Applied Soft Computing*, vol. 96, pp. 1–11, 2020.

[17] P. Mishra and S. K. Panda, "Zero-day attack detection and simulation through deep learning techniques," in *Proceedings of the IEEE International Conference on Intelligent Systems and Advanced Computing*, 2022, pp. 1–6.

[18] L. Manjula and E. Saravanakumar, "Zero-day attack detection in cloud security: A survey of machine learning and deep learning approaches," in *Proceedings of the IEEE International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2025, pp. 2055–2062.

[19] S. I. B. Vasquez, P. A. H. Monckton, D. I. L. MuÃ±oz, and H. Allende, "Zero-day threat mitigation via deep learning in cloud environments," *Applied Sciences*, vol. 15, no. 14, p. 7885, 2025.

[20] A. Naithani, S. N. Singh, and P. K. Sharma, "Zero-day vigilance: Proposing a machine learning framework for DDoS detection in cloud computing," in *Proceedings of the IEEE International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*, 2024, pp. 818–824.

Towards Reliable Recognition of Concurrent Abnormal Patterns in Control Charts Using Multi-Label Deep Learning

Mohammed Modar[1], Abdelilah Ganmati[2], Omar El Farissi[3]
Industrial Engineering Department, National School of Applied Sciences, Agadir, Morocco[1]
Laboratory of Computer Systems and Vision (LABSIV), Faculty of Sciences, Agadir, Morocco[2]
Mechanical Engineering Department, National School of Applied Sciences, Agadir, Morocco[3]

***Abstract*—Control charts do more than raise an alarm: their shapes can give an early indication of what has changed in a process. This study considers the case in which one chart window contains more than one abnormal behavior. The observed sequence is then a mixture rather than a pure pattern. We formulate this problem directly as multi-label classification. A one-dimensional CNN receives a raw-scale window of 32 observations and predicts the active elementary labels. The controlled protocol contains twelve scenarios: normal behavior, six single abnormal patterns, and five selected concurrent patterns. Raw-scale input is retained because shift patterns depend partly on level information that may be weakened by window-wise normalization. The retained training setup gives additional exposure to difficult shift and trend cases, while validation and testing remain balanced. Across five repeated trainings, the model achieved 96.11% exact match accuracy, 96.41% precision, 96.46% recall, 96.44% F1-score, and 1.04% Hamming loss. The 95% confidence interval for exact match was 96.05–96.17%. Additional analyses show stable performance around a decision threshold of 0.5, strong cyclic and systematic recognition, and lower performance for short shift cases. The results support direct multi-label CNN recognition for the selected protocol. Broader shift-containing mixtures, more complex combinations, varying noise conditions, and real industrial validation remain outside the scope of the present controlled study.**

Keywords*—Control chart pattern recognition; concurrent patterns; statistical process control; multi-label classification; convolutional neural network; process monitoring

I. Introduction

Control charts remain one of the basic tools of statistical process control. When a process is stable, the plotted observations fluctuate around a reference level without a persistent structure. When an assignable cause appears, the chart may drift, shift to another level, oscillate, or alternate in a systematic way [1]–[5]. These shapes are useful because they can provide an early diagnostic indication. A trend may be associated with progressive tool wear or a gradual operating change, while a shift may reflect an abrupt change in machine settings, material, or method. Cyclic and systematic patterns may indicate periodic influences or alternating production conditions [1], [5], [16], [23].

The goal of control chart pattern recognition (CCPR) is to identify the abnormal behavior visible in a chart window. Earlier studies used neural networks, statistical features, support vector machines, and optimization-based classifiers [6]–[13]. More recent work has used convolutional models, recurrent networks, feature fusion, transfer learning, and attention-based architectures [23], [29], [36]–[39].

Much of the CCPR literature assumes that one window contains one dominant abnormal pattern. This assumption is useful for basic classification, but it is restrictive in practice. Two assignable causes may act at the same time, so a chart can drift while it oscillates, or show a systematic alternation while another change is developing.

Concurrent CCPR addresses this case. It is difficult because the components are mixed in the same short and noisy sequence. A local increase may be random variation, the start of a trend, or part of a shift. A periodic movement may also be partly hidden by noise. The model should therefore indicate which elementary components are present rather than only select the nearest exclusive class.

Several concurrent CCPR studies transform the signal before classification. Wavelet decomposition, independent component analysis, singular spectrum analysis, blind source separation, learning vector quantization, support vector machines, random forests, and extreme learning machines have been used for this purpose [14]–[22]. These studies show that concurrent CCPR is not a minor extension of ordinary pattern recognition. They also show that the representation matters: when decomposition is inaccurate, its error is passed to the classifier.

Another issue is the representation of the target. A common approach is to create a separate class for each mixture. This is manageable when the number of combinations is small, but every new mixture requires a new class and the elementary diagnostic components remain hidden inside the class name. In this work, each elementary behavior has its own output. A window containing an upward trend and a cyclic movement should activate both labels. This description is closer to the way an engineer reads the chart.

The present study therefore focuses on a direct multi-label formulation. We use a compact 1D CNN, independent sigmoid outputs, and multi-label evaluation metrics. We keep the input on its original scale because shifts are partly defined by a level change. The main protocol contains twelve scenarios: normal behavior, six single abnormal patterns, and five concurrent cases. This set is deliberately focused. It is large enough to test direct concurrent recognition while remaining clear enough for

pattern-level error analysis.

We also compare a small number of reasonable training variants under the same protocol. The aim is not to claim a universal architecture or an exhaustive hyperparameter search. It is to establish a transparent reference configuration and examine its stability, label-wise behavior, threshold sensitivity, computational cost, and practical scope.

The main contributions are as follows:

- Concurrent CCPR is formulated as a direct multi-label problem so that the elementary abnormal components remain visible in the output.
- A twelve-scenario protocol is defined with explicit encoding for normal, single abnormal, and selected concurrent cases.
- A compact 1D CNN is evaluated on raw-scale windows using full-vector and label-wise metrics across repeated runs.
- Additional analyses examine uncertainty, threshold sensitivity, label-wise errors, computational cost, and the practical limits of the focused experimental scope.

The remainder of the study is organized as follows. Section II reviews related CCPR and multi-label studies. Section III presents the methodology. Section IV describes the experimental setup. Section V reports and discusses the results. Section VI concludes the study.

II. Literature Review

A. Single Control Chart Pattern Recognition

Early CCPR studies generally considered one abnormal pattern at a time. Neural networks were used to map a chart window to a pattern class [6], [7], while other studies used statistical or shape-based features to improve robustness to noise [10]–[12]. Support vector machines and optimization-based models were later adopted for nonlinear decision boundaries [13]. More recent work has introduced 1D CNNs, feature fusion, transfer learning, recurrent networks, and methods designed for imbalanced data or small shifts [29], [36]–[39].

The single-pattern setting remains useful because it isolates the shapes of trends, shifts, cycles, and systematic alternations. Its main limitation is that each window is forced into one class even when more than one source is active.

B. Concurrent Pattern Recognition

Concurrent CCPR was developed for mixed abnormal structures. Table I summarizes representative approaches in this area. Many methods first transform or decompose the signal and then classify the resulting representation. Wavelet methods use time-frequency information [14], [17]; ICA and blind source separation attempt to recover hidden components [15], [20], [21]; and singular spectrum analysis has been combined with SVM, LVQ, and random forest classifiers [16], [18]. Other studies have used ESMD, extreme learning machines, and PCA-weighted strategies [19], [22].

These methods show two recurring issues. First, overlapping patterns are harder to separate than single patterns. Second, performance depends not only on the classifier but also on the representation produced before classification. A decomposition stage can be useful, but it also introduces another source of error.

The target representation is equally important. Composite-class methods assign one class to each predefined mixture. A label-wise formulation instead keeps the same elementary labels across single and concurrent windows. For example, the cyclic label has the same meaning whether it appears alone, with a trend, or with another component. This representation supports a more direct diagnostic reading, although it does not by itself identify the physical root cause.

C. Deep Learning and Multi-Label Learning

Deep learning is now widely used for time-series classification and process monitoring [25]–[27]. For short fixed-length windows, a 1D CNN is a natural baseline because its filters can react to local rises, drops, transitions, oscillations, and alternating segments [28], [29], [35], [36]. Recurrent, hybrid CNN–LSTM, residual, and attention-based models are useful alternatives when longer temporal dependencies or variable-length sequences are central [24], [38], [39].

The output layer is central to the present problem. Softmax is appropriate when only one class can be active. Concurrent patterns require independent outputs because a cyclic component can coexist with a trend or a shift. We therefore use one sigmoid output per label and optimize binary cross-entropy, following standard multi-label learning practice [32]–[34].

The evaluation must also reflect partial correctness. Exact match requires the entire label vector to be correct, whereas Hamming loss counts individual label errors. Precision, recall, and F1-score show whether the model tends to create false activations or miss active components. The present work combines concurrent CCPR with this label-wise output and evaluation view.

III. Proposed Methodology

Fig. 1 shows the workflow we used in the experiments. We start from basic control chart components and use them to create single-pattern windows and selected concurrent windows. After that, each window is encoded as a vector of seven binary labels and passed to a one-dimensional CNN. We then read the prediction as a multi-label output, not as a single class.

We kept the method simple for a reason. If the generator is not clear, it becomes difficult to know why the model succeeds or fails. Also, the target must follow the physical meaning of the signal. If two components are used to generate the chart, two labels should be active. Finally, the evaluation should not hide partial mistakes. Detecting one component and missing the second is not the same error as detecting nothing.

This also explains why we do not try to solve all possible CCPR cases at once. The number of possible concurrent combinations grows quickly when more basic patterns are added. A very large set would look more complete, but it would be harder to understand. Here, the aim is to keep enough cases to test the concurrent idea and still be able to read the errors pattern by pattern.

TABLE I. REPRESENTATIVE CHOICES IN CONCURRENT CCPR LITERATURE

Study	Main strategy	Output view	Main relevance to the present work
Al-Assaf [14]	Wavelet-based multiresolution analysis	Pattern class	Shows that time-frequency representation can support concurrent CCPR.
Lu et al. [15], [21]	ICA combined with SVM	Separated patterns	Uses decomposition before classification; performance depends on separation quality.
Xie et al. [16]	SSA combined with SVM	Composite class	Represents concurrent windows through extracted temporal components.
Du et al. [17]	Wavelet decomposition and multiclass SVM	Composite class	Treats each concurrent case as a class after transformed feature extraction.
Gu et al. [18]	SSA and learning vector quantization	Composite class	Uses decomposition and prototype-based recognition for mixed patterns.
Yang et al. [19]	ESMD and extreme learning machine	Components	Addresses identification and quantification of concurrent patterns.
Pelegrina et al. [20]	Blind source separation and feature extraction	Separated sources	Highlights the importance and sensitivity of signal separation.
Cheng et al. [23]	Multi-label 1D CNN	Active labels	Provides the closest direct multi-label CNN reference for concurrent CCPR.

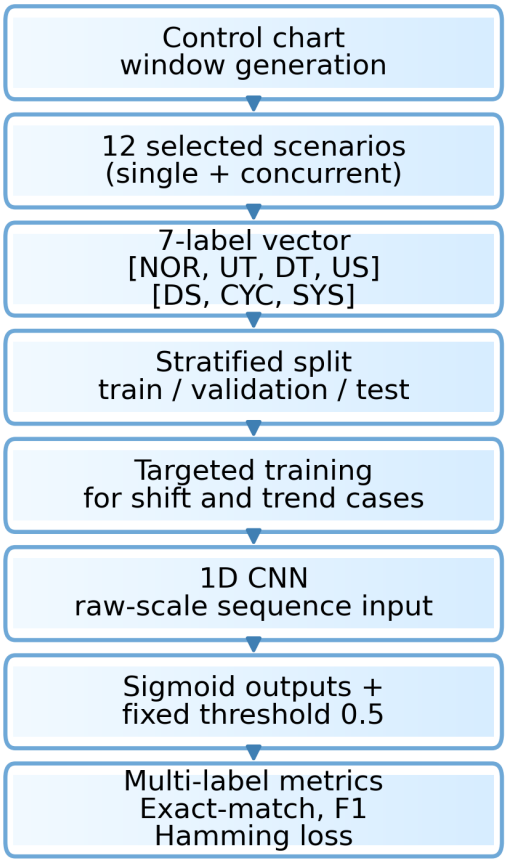


Fig. 1. Experimental workflow used in the proposed study.

We use the same logic for the model. We do not add a decomposition block or a long list of handcrafted features. Such elements may be useful later, but they would make it harder to see the effect of the multi-label formulation itself. We therefore use the present design as a direct reference point: raw window, CNN, sigmoid labels, and multi-label metrics.

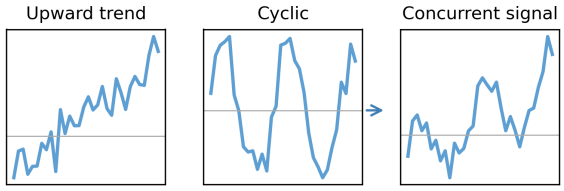


Fig. 2. Illustration of the transition from single abnormal patterns to a concurrent observed signal.

A. From Single Pattern to Concurrent Patterns

Going from single patterns to concurrent patterns changes the problem quite strongly. In the single case, the model searches for one dominant abnormal source. In the concurrent case, two sources may be present at the same time. The target is therefore not a class name anymore; it is the set of active components.

Fig. 2 gives a small example. One signal has a trend, another has a cyclic movement, and the final signal is the visible mixture. A classical multiclass model would need a special class for this mixture. With the multi-label view, the two components remain visible. This is more useful for diagnosis, because the output says what is inside the mixture.

B. Problem Formulation

Let a control chart window be represented as a one-dimensional sequence:

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T, \tag{1}$$

where, $n = 32$ in our study. We aim to learn a function f_θ that maps the input window to a vector of label probabilities:

$$\mathbf{p} = f_\theta(\mathbf{x}) = [p_{NOR}, p_{UT}, p_{DT}, p_{US}, p_{DS}, p_{CYC}, p_{SYS}], \tag{2}$$

where, each element belongs to $[0, 1]$. The corresponding ground-truth vector is:

$$\mathbf{y} = [y_1, y_2, \ldots, y_7], \quad y_k \in \{0, 1\}. \tag{3}$$

We represent a concurrent pattern by several active labels in $\mathbf{y}$. For example, an upward trend with a cyclic component has $y_{UT} = 1$ and $y_{CYC} = 1$.

We do not treat this as a softmax problem. Softmax makes the classes compete, which is correct only when one class excludes the others. Here, one label does not exclude another one. A cyclic component can appear with a trend. For that reason, each output neuron uses a sigmoid activation.

We keep the normal label as an explicit output. We do not combine it with abnormal labels in our generated scenarios. A normal window means that no deterministic abnormal component has been added. This gives the CNN a reference behavior, against which the abnormal components are learned.

We make this point to avoid a small ambiguity. Random noise is present in all windows, but the label NOR does not mean 'noise'. It means that none of the deterministic abnormal components of the protocol is present.

C. Pattern Set and Label Encoding

We use twelve scenarios. Seven are single-label cases: normal, upward trend, downward trend, upward shift, downward shift, cyclic, and systematic. The remaining five are concurrent cases built from trend, cyclic, and systematic components. Shift-containing mixtures are not part of the present concurrent set. This choice is intentional rather than a conceptual exclusion. In a short window of 32 observations, the visible form of a shift depends jointly on its magnitude, change-point location, and the number of post-change observations. When a shift is superposed on another component, the relative amplitudes and temporal alignment introduce additional factors that require a dedicated generation and training protocol. Adding these cases to the present experiment would therefore change the controlled scope rather than simply add another class.

The scenario set is consequently a focused reference rather than an exhaustive catalogue of industrial mixtures. It was selected to examine the direct multi-label formulation under clearly defined and interpretable conditions. We do not claim that the resulting model covers every possible combination. We also do not use scenario names as model outputs. Instead, we use elementary diagnostic components. For example, *upward trend + cyclic* activates two labels, while *cyclic* activates one. Table II separates the scenario used to generate the window from the labels predicted by the model. This distinction allows a concurrent window to be described through more than one component.

Fig. 3 illustrates representative windows for the twelve scenarios used in the controlled protocol.

D. Data Generation Protocol

Each window is generated by adding a deterministic pattern component to random noise:

TABLE II. PATTERN SCENARIOS AND ACTIVE LABEL VECTORS

Scenario	Active labels	Label-vector interpretation
Normal	NOR	In-control behavior.
Upward trend	UT	Increasing slope.
Downward trend	DT	Decreasing slope.
Upward shift	US	Positive level change.
Downward shift	DS	Negative level change.
Cyclic	CYC	Oscillatory component.
Systematic	SYS	Alternating high-low structure.
Upward trend + cyclic	UT, CYC	Trend and oscillation.
Downward trend + cyclic	DT, CYC	Negative trend and oscillation.
Upward trend + systematic	UT, SYS	Trend with alternation.
Downward trend + systematic	DT, SYS	Negative trend with alternation.
Systematic + cyclic	SYS, CYC	Alternation and oscillation.

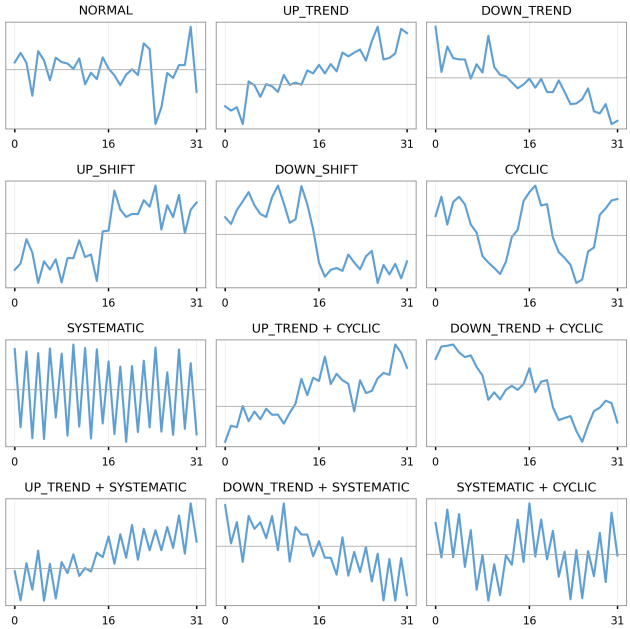


Fig. 3. Examples of the twelve generated control chart scenarios.

$$x_t = \epsilon_t + D_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \quad t = 1, \ldots, n. \tag{4}$$

We define the term D_t according to the active labels. The deterministic components are summarized in Table III. In a concurrent case, the deterministic components are added first, and noise is then added. In the main protocol, the Gaussian noise level is fixed at $\sigma = 1$ for all training, validation, and test windows. This controlled setting isolates the effect of the multi-label formulation and keeps the comparison between model configurations consistent. A trend creates drift, a shift creates a level change after a point, a cyclic term creates oscillation, and the systematic term gives a high-low alternation. We therefore know the labels from the generator; we do not assign them by visual judgment.

TABLE III. DETERMINISTIC COMPONENTS USED IN THE GENERATOR

Component	Expression	Role
Trend	$\pm gt$	Models progressive drift.
Shift	$\pm s\mathbb{1}(t > \tau)$	Models a sudden level change.
Cyclic	$a\sin(2\pi t/T + \phi)$	Models periodic behavior.
Systematic	$d(-1)^t$	Models alternating high-low observations.
Concurrent case	$D_t = D_t^{(1)} + D_t^{(2)}$	Adds active components in one window.

We keep the additive form plain by design. It keeps a visible link between the signal and the label vector. If the model misses a label, we know which component was missed. This is useful in a first controlled study.

In the main experiment, we keep the signal on its original scale. We do not apply window-wise z-score normalization. The reason is simple: a shift is partly a level change, and this level information can be weakened by normalizing every window separately.

We do not mean that normalization is never useful in CCPR. We only mean that, for the present target labels, preserving the vertical level is more consistent. This is especially true for upward and downward shifts.

This also makes our error discussion easier. If a shift is confused with a trend, we can still look at the original level and slope of the window. With independent normalization, part of this information has already been changed.

We also use the generator to improve reproducibility. Concurrent-CCPR papers are difficult to compare when equations, parameter ranges, or window sizes are not given in enough detail. Here, training, validation, and test windows follow the same definitions, but they are generated with different random seeds. The test data remain unseen, although they come from the same controlled process.

E. CNN Architecture

The CNN receives a raw sequence of 32 observations. Fig. 4 and Table IV summarize the retained architecture, which contains two one-dimensional convolutional layers, max pooling, dropout, one dense layer, and seven sigmoid outputs. No handcrafted descriptors or decomposition stage is added before the network.

The compact architecture is intentional. With only 32 input points, two convolutional stages are sufficient to build local features at two levels without introducing a large parameter count. The first layer detects short transitions and repeated local movements, while the second combines them into broader pattern evidence. A deeper residual network, a CNN–LSTM hybrid, or an attention mechanism may be useful for longer, variable-length, or more irregular signals [38], [39]; however, such models would add capacity and tuning choices that are outside the focused comparison made here.

The retained model has 15,687 trainable parameters and approximately 54,680 multiply-accumulate operations per sample. This small computational footprint supports its use as a reproducible reference architecture.

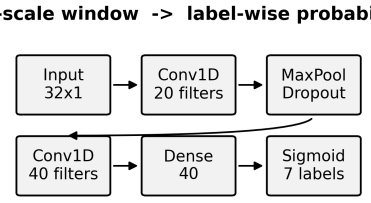


Fig. 4. CNN architecture used for multi-label recognition of control chart windows.

TABLE IV. CNN CONFIGURATION USED IN THE RETAINED EXPERIMENT

Element	Setting
Input size	32×1 raw-scale signal.
Convolution 1	20 filters, kernel size 5, ReLU.
Pooling/dropout	Max pooling followed by 12% dropout.
Convolution 2	40 filters, kernel size 3, ReLU.
Dense layer	40 units, ReLU, 12% dropout.
Output layer	Seven sigmoid neurons.
Loss	Binary cross-entropy.
Optimizer	Adam, learning rate 0.001.
Decision rule	Fixed threshold of 0.5.
Trainable parameters	15,687.
Approximate MACs	54,680 per sample.

F. Learning Objective and Decision Rule

We train the CNN using binary cross-entropy averaged over the $K = 7$ labels:

$$\mathcal{L}(\mathbf{y}, \mathbf{p}) = -\frac{1}{K}\sum_{k=1}^{K}\left[y_k \log(p_k) + (1 - y_k)\log(1 - p_k)\right]. \quad (5)$$

We interpret the sigmoid output p_k as the probability that label k is active. At inference time, the decision rule is:

$$\hat{y}_k = \begin{cases} 1, & p_k \geq 0.5, \\ 0, & p_k < 0.5. \end{cases} \quad (6)$$

We fix the threshold before testing. A label is active when its sigmoid value is at least 0.5. Label-specific thresholds could improve particular cases, but they would add another tuning stage. We therefore keep one common threshold and examine its sensitivity separately in Section V using saved prediction probabilities.

G. Evaluation Metrics

We choose the metrics according to the output. In many concurrent-CCPR studies, recognition accuracy or correct classification rate is enough because the output is one composite class [15]–[22]. Here, the output is a vector. We therefore need to look at both the full vector and the individual label decisions.

We use exact match as the strict measure: the complete vector must be correct. We use Hamming loss as a softer measure; it counts the wrong labels one by one. Precision tells

TABLE V. MULTI-LABEL METRICS RETAINED FOR EVALUATION

Metric	Expression	Interpretation
Exact match	$\frac{1}{N}\sum_i \mathbb{1}(\mathbf{y}_i = \hat{\mathbf{y}}_i)$	Strict full-vector correctness.
Precision	$TP/(TP+FP)$	Penalizes false label activations.
Recall	$TP/(TP+FN)$	Penalizes missed active labels.
F1-score	$2PR/(P+R)$	Balances precision and recall.
Hamming loss	$\frac{1}{NK}\sum_{i,k} \mathbb{1}(y_{ik} \neq \hat{y}_{ik})$	Label-wise error rate.

us whether the model activates labels too easily. Recall tells us whether it misses active labels. F1-score gives a balance. Table V summarizes these metrics and their interpretation [32]–[34].

H. Rationale for the Main Design Choices

We fixed a few choices before the final comparison. The first one was the window length. The window should be long enough to show a trend, a shift, or a cycle. It should also be short enough to keep the detection useful. We used 32 observations.

Our second choice was the raw input. Many concurrent-CCPR methods transform the signal before classification. This can help, but it also adds parameters and possible errors. In this study, we ask a narrower question: what can be achieved if the CNN sees only the observed chart window? This gives us a simple reference for later comparisons.

Our third choice was about balance. We gave each scenario the same number of test windows. The label distribution is not perfectly balanced, because some labels appear in more than one scenario. This is expected. We give the retained training setup more examples of difficult shift and trend cases, but we keep validation and testing balanced.

Our fourth choice concerned the metrics. Exact match tells us whether the whole diagnosis is correct, but it does not tell us how large the error is. For that reason, we also report Hamming loss, precision, recall, and F1-score.

Finally, we keep the number of variants limited. This is not a broad hyperparameter search. We compare a few reasonable choices: baseline CNN, two-channel input, weighted loss, shift-extra-only training, and targeted training. We keep the comparison readable.

IV. EXPERIMENTAL SETUP

The data follow the protocol described in Section III, and Table VI summarizes the main experimental settings. For each scenario, 5000 windows are generated for training and 1250 for testing. Ten percent of the training data are reserved for stratified validation. The test set remains balanced across the twelve scenarios and is not used to select the decision threshold.

The CNN is implemented in TensorFlow [30] and trained with Adam, a batch size of 16, and a maximum of 100 epochs. Validation loss is used for early stopping, and the evaluation metrics are computed with scikit-learn [31]. The experiment is repeated five times with different seeds. The retained configuration receives additional training examples for difficult shift and trend cases, but the validation and test sets remain unchanged.

TABLE VI. EXPERIMENTAL SETTINGS

Item	Value
Window size	32 observations.
Pattern scenarios	12 in the main protocol.
Output labels	7: NOR, UT, DT, US, DS, CYC, SYS.
Training samples	5000 windows per scenario before reinforcement.
Test samples	1250 windows per scenario, balanced.
Validation split	10% stratified validation.
Preprocessing	Raw-scale signal; no window-wise z-score.
Gaussian noise	Fixed at $\sigma = 1$ in all data splits.
Decision threshold	0.5 for all labels.
Number of runs	5 repeated trainings.

TABLE VII. OVERALL PERFORMANCE OF THE RETAINED CONFIGURATION

Metric	Mean (%)	Std.	95% CI (%)
Exact match	96.11	0.05	[96.05, 96.17]
Precision	96.41	0.05	[96.35, 96.47]
Recall	96.46	0.07	[96.37, 96.55]
F1-score	96.44	0.06	[96.37, 96.51]
Hamming loss	1.04	0.02	[1.02, 1.06]

No handcrafted descriptors, wavelet transform, SSA, ICA, or blind-source-separation stage is used. The main purpose is to examine how far a direct multi-label CNN can go with the observed raw-scale window. This design leaves fewer intermediate choices and makes the reference experiment easier to reproduce.

Reviewer-directed analyses were performed without changing the main twelve-scenario protocol. Saved test probabilities were used for threshold sensitivity and label-wise evaluation, and a matched five-run baseline was trained for paired statistical tests. These additions strengthen the analysis while preserving the original data-generation and evaluation scope.

V. RESULTS AND DISCUSSION

A. Overall Multi-Label Performance

Table VII reports the main five-run result for the retained configuration. Mean exact match is 96.11%, F1-score is 96.44%, and Hamming loss is 1.04%. The narrow confidence intervals indicate that the result is not driven by one favorable initialization.

Hamming loss gives a label-level view of the error. A value of 1.04% means that only a small fraction of individual label decisions are wrong, even when a full label vector is counted as incorrect by exact match. A separate five-run verification performed for the reviewer analyses produced 96.05% exact match and 96.39% F1-score, which is close to the original result and supports reproducibility.

B. Label-Wise and Pattern-Level Results

Fig. 5 shows precision, recall, and F1-score for each output label using the ensemble mean probability across the five verification runs. Cyclic and systematic labels are the strongest,

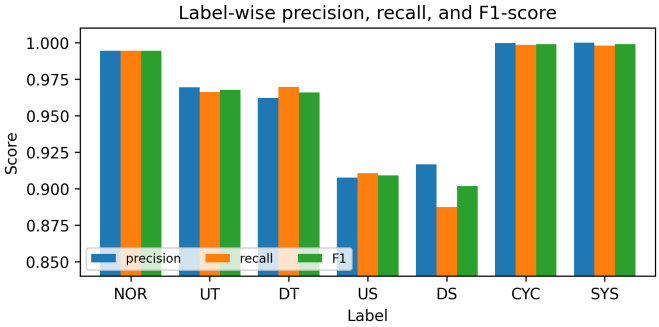


Fig. 5. Label-wise precision, recall, and F1-score.

TABLE VIII. PATTERN-LEVEL EXACT MATCH PERFORMANCE

Scenario	Exact match (%)	Test windows
Downward shift	88.16	1250
Upward trend	90.00	1250
Downward trend	91.04	1250
Upward shift	91.36	1250
Systematic	99.36	1250
Upward trend + cyclic	99.36	1250
Cyclic	99.44	1250
Normal	99.60	1250
Downward trend + cyclic	99.68	1250
Downward trend + systematic	99.68	1250
Upward trend + systematic	99.68	1250
Systematic + cyclic	99.92	1250

TABLE IX. COMPARISON OF EXPLORED CONFIGURATIONS

Configuration	Exact match (%)	F1-score (%)	Hamming loss (%)
Baseline CNN	96.08	96.40	1.04
Two-channel input variant	95.83	96.17	1.11
Targeted-training CNN	96.11	96.44	1.04
Weighted-loss variant	95.95	96.35	1.06
Shift-extra-only variant	95.90	96.24	1.09

with F1-scores of 99.90% and 99.89%, respectively. Upward and downward trend remain above 96.5% F1. Shift labels are weaker: upward shift reaches 90.89% F1 and downward shift 90.16%. The downward-shift false-negative rate is 11.28%, compared with 8.96% for upward shift.

Table VIII and Fig. 6 give the original scenario-level exact match rates. The selected concurrent cases are recognized well, whereas the weaker cases are single shifts and trends. This difference is plausible in a 32-point window. A weak trend may resemble random drift, while a small or late shift leaves only a limited number of post-change observations. In contrast, a cycle or systematic alternation creates repeated evidence across the window.

C. Comparison of Explored Configurations

Table IX summarizes the configurations examined before the retained setup was selected. The differences between the strongest variants are small. The retained targeted-training model is therefore treated as a balanced configuration, not as a universally superior architecture.

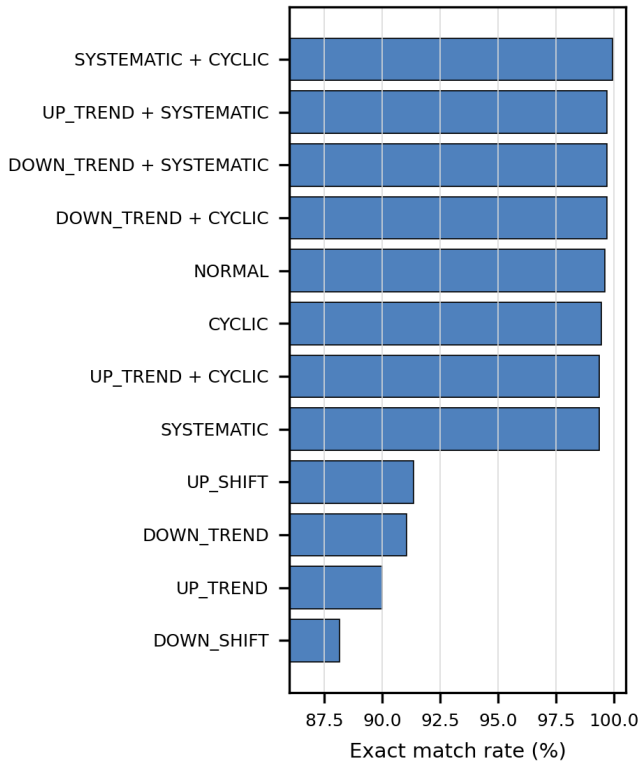


Fig. 6. Pattern-level exact match rate for the twelve control chart scenarios.

To test whether the small baseline difference was meaningful, a new matched five-run comparison was performed using the same seeds. The repeat means were 95.99% versus 96.05% for exact match and 96.33% versus 96.39% for F1-score. Paired t-tests gave $p = 0.558$ for exact match and $p = 0.653$ for F1-score; Wilcoxon tests led to the same conclusion. The difference is not statistically significant. We therefore retain the targeted-training setup because it gave the most balanced behavior across the difficult cases, while avoiding a claim of statistical superiority.

D. Decision-Threshold Sensitivity

Fig. 7 evaluates one common threshold from 0.30 to 0.70. Lower thresholds increase recall and give the highest F1-score at 0.30 (96.96%), but strict exact match falls to 93.98% because more false labels are activated. The threshold of 0.5 gives the highest exact match (96.05%) and the lowest Hamming loss (1.05%) in the repeated analysis. It is therefore a reasonable common operating point for the present objective.

E. Comparison with Prior Concurrent CCPR Studies

The closest published study is the multi-label CNN of Cheng et al. [23]. Table X provides a direct numerical context using their first experiment, which also used 32-point windows, seven labels, 5000 training samples per scenario, and 1250 test samples per scenario. Their protocol contains more concurrent scenarios, whereas the present protocol concentrates on five selected mixtures and preserves raw level information for shifts. The comparison is therefore informative but not a controlled head-to-head benchmark.

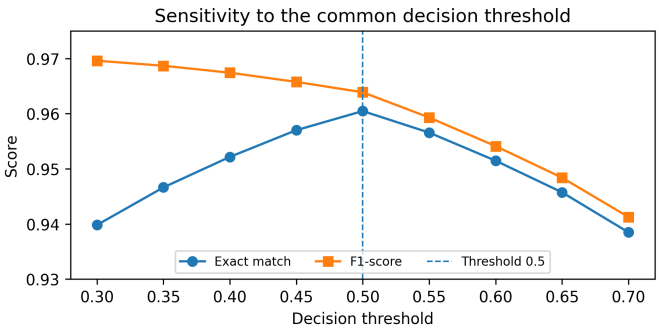


Fig. 7. Sensitivity of exact match and F1-score to the common decision threshold.

TABLE X. CONTEXTUAL COMPARISON WITH THE CLOSEST MULTI-LABEL CNN STUDY

Study	Scenarios	Window	Exact match (%)	F1 (%)	Hamming loss (%)
Cheng et al. [23], Exp. 1	16	32	92.56	95.68	1.98
Present study	12	32	96.11	96.44	1.04

The numerical values should not be read as proof of general superiority because the pattern generators and scenario sets differ. Cheng et al. also included a real-world monitoring demonstration, which is an important strength of their study. The contribution here is narrower: a compact raw-scale reference, explicit repeated-run stability, focused analysis of difficult shift and trend labels, and a transparent evaluation of the selected protocol. Earlier SSA–SVM, ICA–SVM, wavelet, and ELM studies remain relevant, but their composite-class outputs and different generators make direct metric comparison unreliable [15]–[22].

F. Computational Cost

The retained network is lightweight: it has 15,687 trainable parameters, about 54,680 MACs per sample, and a saved model size of approximately 0.22 MB. In the Colab CPU reruns used for the reviewer analysis, mean training time was 582 ± 117 s. Batched inference over the 15,000-window test set took approximately 0.29 s, corresponding to about 0.019 ms per sample. These values are implementation-dependent, but they show that the compact model has modest computational requirements.

G. Limitations and Practical Scope

The main results are obtained from synthetic signals. This controlled setting is useful because the active label vector is known exactly, which is rarely available for concurrent abnormalities in real plant data. However, it does not demonstrate industrial generalization. Sensor effects, autocorrelation, changes in variance, maintenance actions, and sampling changes may alter the observed shapes.

The scenario set is also intentionally limited to five selected concurrent combinations. Shift-containing mixtures, combinations with more than two simultaneous abnormalities, and systematic variation of the noise level were not included in the present protocol. A scientifically consistent extension would require a broader factorial design that controls shift magnitude, change-point position, component amplitudes, temporal alignment, and noise intensity in both training and testing. Adding a small number of post hoc cases would instead evaluate a different problem and would not provide a fair estimate of the model under a properly designed extended protocol. The claims of this study are therefore restricted to the twelve scenarios defined in Section III.

In practical use, the output should be treated as decision support rather than a closed diagnosis. A high probability means that the window resembles a learned component, not that the physical root cause has been proven. Real production validation with expert-confirmed multi-label annotations, broader concurrent combinations, and controlled noise studies remain necessary next steps.

VI. CONCLUSION

This study studied concurrent abnormal pattern recognition in control charts using a direct multi-label CNN. Each window is represented by seven elementary labels, so more than one abnormal behavior can be active without creating a separate composite class for every mixture.

Under the controlled twelve-scenario protocol, the retained model achieved 96.11% exact match accuracy, 96.44% F1-score, and 1.04% Hamming loss across five repeated trainings. Confidence intervals were narrow, and threshold analysis supported 0.5 as a stable common operating point when strict full-vector correctness is prioritized. Label-wise analysis showed very strong cyclic and systematic recognition, while short shifts and trends remained more difficult.

The method should therefore be understood as a compact and interpretable reference for the selected concurrent cases, not as a complete solution for every possible mixture. Shift-containing concurrent patterns, combinations with more than two abnormalities, varying noise conditions, and real industrial data require a dedicated extension of the generation and training protocol. These directions are important future work, but they do not change the main conclusion obtained for the controlled twelve-scenario setting.

DISCLOSURE AND CONFLICT OF INTEREST

We declare no conflict of interest. No financial or personal relationship influenced the design, analysis, or reporting of this work.

DECLARATION ON GENERATIVE AI

We used language and formatting tools only to support manuscript preparation. We reviewed the manuscript and remain responsible for its scientific content.

REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control*, 8th ed. Hoboken, NJ, USA: Wiley, 2020.

[2] Western Electric Company, *Statistical Quality Control Handbook*. Indianapolis, IN, USA: Western Electric Company, 1956.

[3] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*. New York, NY, USA: D. Van Nostrand, 1931.

[4] L. S. Nelson, "The Shewhart control chart tests for special causes," *Journal of Quality Technology*, vol. 16, no. 4, pp. 237-239, 1984.

[5] L. S. Nelson, "Interpreting Shewhart control charts," *Journal of Quality Technology*, vol. 17, no. 2, pp. 114-116, 1985.

[6] C. S. Cheng, "A neural network approach for the analysis of control chart patterns," *International Journal of Production Research*, vol. 35, no. 3, pp. 667-697, 1997.

[7] R. S. Guh and J. D. T. Tannock, "Recognition of control chart concurrent patterns using a neural network approach," *International Journal of Production Research*, vol. 37, no. 8, pp. 1743-1765, 1999.

[8] F. Zorriassatine and J. D. T. Tannock, "A review of neural networks for statistical process control," *Journal of Intelligent Manufacturing*, vol. 9, pp. 209-224, 1998.

[9] W. Hachicha and A. Ghorbel, "A survey of control-chart pattern-recognition literature (1991-2010) based on a new conceptual classification scheme," *Computers & Industrial Engineering*, vol. 63, no. 1, pp. 204-222, 2012.

[10] A. Hassan, M. S. N. Baksh, A. M. Shaharoun, and H. Jamaluddin, "Improved SPC chart pattern recognition using statistical features," *International Journal of Production Research*, vol. 41, no. 7, pp. 1587-1603, 2003.

[11] D. T. Pham and M. A. Wani, "Feature-based control chart pattern recognition," *International Journal of Production Research*, vol. 35, no. 7, pp. 1875-1890, 1997.

[12] S. K. Gauri and S. Chakraborty, "Recognition of control chart patterns using improved selection of features," *Computers & Industrial Engineering*, vol. 56, no. 4, pp. 1577-1588, 2009.

[13] V. Ranaee, A. Ebrahimzadeh, and R. Ghaderi, "Application of the PSO-SVM model for recognition of control chart patterns," *ISA Transactions*, vol. 49, no. 4, pp. 577-586, 2010.

[14] Y. Al-Assaf, "Multi-resolution wavelets analysis approach for the recognition of concurrent control chart patterns," *Quality Engineering*, vol. 17, no. 1, pp. 11-21, 2005.

[15] C. J. Lu, Y. E. Shao, and C. C. Li, "Recognition of concurrent control chart patterns by integrating ICA and SVM," *Applied Mathematics & Information Sciences*, vol. 8, no. 2, pp. 681-689, 2014.

[16] L. Xie, N. Gu, D. Li, Z. Cao, M. Tan, and S. Nahavandi, "Concurrent control chart patterns recognition with singular spectrum analysis and support vector machine," *Computers & Industrial Engineering*, vol. 64, no. 1, pp. 280-289, 2013.

[17] S. Du, D. Huang, and J. Lv, "Recognition of concurrent control chart patterns using wavelet transform decomposition and multiclass support vector machines," *Computers & Industrial Engineering*, vol. 66, no. 4, pp. 683-695, 2013.

[18] N. Gu, Z. Cao, L. Xie, D. Creighton, M. Tan, and S. Nahavandi, "Identification of concurrent control chart patterns with singular spectrum analysis and learning vector quantization," *Journal of Intelligent Manufacturing*, vol. 24, pp. 1241-1252, 2013.

[19] W. A. Yang, W. Zhou, W. Liao, and Y. Gou, "Identification and quantification of concurrent control chart patterns using extreme-point symmetric mode decomposition and extreme learning machines," *Neurocomputing*, vol. 147, pp. 260-270, 2015.

[20] G. D. Pelegrina, L. T. Duarte, and C. Jutten, "Blind source separation and feature extraction in concurrent control charts pattern recognition," *Computers & Industrial Engineering*, vol. 92, pp. 105-114, 2016.

[21] C. J. Lu, Y. E. Shao, and P. H. Li, "Mixture control chart patterns recognition using independent component analysis and support vector machine," *Neurocomputing*, vol. 74, no. 11, pp. 1908-1914, 2011.

[22] Z. Zhou, P. Jiang, and X. Wang, "PCA-WA based approach for concurrent control chart pattern recognition," *Computers & Industrial Engineering*, vol. 104, pp. 1-12, 2017.

[23] C. S. Cheng, P. W. Chen, and Y. Ho, "Control chart concurrent pattern classification using multi-label convolutional neural networks," *Applied Sciences*, vol. 12, no. 2, Art. 787, 2022.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[27] F. Chollet, *Deep Learning with Python*. Shelter Island, NY, USA: Manning, 2018.

[28] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. IJCNN*, 2017, pp. 1578-1585.

[29] J. W. Huang, P. J. Lee, and B. P. Jaysawal, "Multiscale control chart pattern recognition using histogram-based representation of value and zero-crossing rate," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 10, pp. 10751-10761, 2022.

[30] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[32] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1-13, 2007.

[33] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819-1837, 2014.

[34] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Cham, Switzerland: Springer, 2016.

[35] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Engineering and Technology*, 2017, pp. 1-6.

[36] L. Xue, H. Wu, H. Zheng, and Z. He, "Control chart pattern recognition for imbalanced data based on multi-feature fusion using convolutional neural network," *Computers & Industrial Engineering*, vol. 182, Art. 109410, 2023.

[37] Y. Li, W. Dai, and Y. He, "Control chart pattern recognition under small shifts based on multi-scale weighted ordinal pattern and ensemble classifier," *Computers & Industrial Engineering*, vol. 189, Art. 109940, 2024.

[38] T. Zan, J. Chen, Z. Sun, Z. Liu, M. Wang, X. Gao, and P. Gao, "Pattern recognition of control chart with variable chain length based on recurrent neural network," *Advanced Manufacturing*, vol. 1, Art. 0003, 2024.

[39] T. Zan, X. Jia, X. Guo, M. Wang, X. Gao, et al., "Research on variable-length control chart pattern recognition based on sliding window method and SECNN-BiLSTM," *Scientific Reports*, vol. 15, Art. 5921, 2025.

Deep Learning for the Classification of Kidney Diseases in Medical Images Using ResNet-50 and Grad-CAM

Laberiano Andrade-Arenas[1], Inooc Rubio Paucar[2], Cesar Yactayo-Arias[3]
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú[1]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[2]
Departamento de Estudios Generales, Universidad Continental, Lima, Perú[3]

***Abstract*—Renal pathology represents a diverse set of diseases that present significant clinical relevance. Included among the various types of renal pathologies are renal stones, cysts, and renal malignancies, all of which require diagnosis and therapy to prevent progression of the disease process. The current research study was performed to create and validate a classification model based on deep learning using a convolutional neural networks (CNN) architecture, namely a 50-layer Residual Network (ResNet-50) using Gradient-weighted Class Activation Mapping (Grad-CAM), to provide improved automatic detection of renal pathology from medical images and improve the interpretability of those medical images. During the study, the Explainable Deep Learning Pipeline (X-DLP) paradigm was followed, which provides a structured methodology to perform research with the use of deep learning in medical imaging. The X-DLP structures the research process into a series of phases, including Data acquisition and curation, Preprocessing and Augmentation, Model Creation via Transfer Learning, and lastly, Interpretability and Visualization.The results obtained show that the proposed model performs consistently well across different evaluation metrics. The Precision–Recall curve, with a PR-AUC close to 0.89, suggests that the model is effective at identifying positive cases even when the data are imbalanced. In addition, the F1-score reaches a peak of around 0.835 at a threshold near 0.45, indicating a good trade-off between precision and recall. From another perspective, the evaluation using Youden's criterion reveals sensitivity and specificity values close to 0.80, which supports the model's ability to distinguish between classes with reasonable accuracy. Moreover, the lift and cumulative gain analysis further highlight its practical usefulness, with a lift of 3.5 in the top 10% and a cumulative gain of 75% when considering 30% of the population. These results indicate that the model can effectively prioritize the most relevant positive cases. Overall, these findings suggest that the model can serve as a valuable support tool in medical diagnosis. By enabling automated classification of renal images and providing visual insights through interpretability techniques, it helps streamline clinical decision-making, reduces reliance on purely manual assessments, and enhances its potential for real-world application.**

***Keywords*—Classification; deep learning; Grad-CAM; medical imaging; ResNet-50**

I. Introduction

The worldwide increase in chronic kidney disease continues to be a major public health issue with millions of individuals affected, along with being a leading factor for mortality and disease morbidity [1]. The World Health Organization (WHO) [2] provides guidance along with the International Society of Nephrology (ISN) [3], that early detection and effective management of chronic kidney disease can help to reduce serious complications, hospitalizations from chronic kidney disease and provide for lower associated healthcare costs, including preventing progression of chronic kidney disease to end-stage renal disease requiring dialysis or kidney transplantation. Unfortunately, the ability to identify renal abnormalities through traditional methods of diagnosing kidney disease is limited due to manual evaluations, variations in interpreting radiographic images and limited access to trained specialists.

Although recent developments have been made in the areas of diagnostic methods and image processing through new imaging techniques; there are still insufficient tools available for early identification of renal abnormalities. The interpretation of medical imaging has a substantial component dependent on the medical expert's level of experience and; as such; can be impacted by inter-observer variability, human flaws and delays in delivery of care, particularly in areas where there are not enough nephrologists or radiologists available [4],[5]. Because of these challenges to effective prevention and identification of individuals with kidney disease, the associated impact of poor clinical outcomes, increased economic burden and decreased quality of life have not been able to be adequately addressed and/or mitigated [6]. As a result, the use of deep learning (DL) methods for assisting in automating and optimising the analysis of images for all types of kidney diseases is beginning to show many significant advances for detection, classification, and follow-up of these diseases [7]. Unfortunately, the most pressing challenge to the acceptance of DL in clinical settings is a lack of interpretability of DL models and; therefore, a lack of trust by physicians and; thus, reluctance to incorporate DL into their day-to-day practice. However, being able to provide visual explanations of DL model decisions is critical to providing transparency, building physician trust and supporting clinical decision making.

This research focuses on developing a deep learning model utilizing the ResNet-50 architecture to automatically classify kidney disease using medical images with the addition of Grad-CAM as an interpretability tool for the results of the model. The output of this system will include both predictions of the potential for abnormalities, as well as visualization of where the model is making those decisions, providing insight to the specialist regarding these areas [8],[9]. This approach follows international standards for quality and safety within

the healthcare profession, facilitating early diagnosis and aiding clinical decision-making for kidney disease, ultimately resulting in improved patient outcomes, while providing a scalable and accessible solution that will help to alleviate the disparity of care in areas where there may be limited numbers of specialists available.

This research project aims to develop a deep learning model based on the ResNet-50 architecture to classify kidney disease through medical images, utilizing interpretability methods like Grad-CAM to assist with clinical decision-making. The system will receive medical images where the physician suspects renal issues and produce probabilities of each type of condition, enabling the physician to predict kidney disease based on the medical images they send.

II. Literature Review

In this section, the theoretical foundations of the research are developed, organized into two main parts. First, the two fundamental variables of the study are addressed: kidney diseases and deep learning, allowing for an understanding of the key concepts that underpin the research. Second, a literature review is presented, in which the contributions of various authors related to the topic are described and analyzed, considering their previous studies and findings. In this way, a solid theoretical framework is built to support the development of the work.

A. Theoretical Base

1) Kidney diseases: Kidney diseases comprise a group of disorders that affect both the structure and function of the kidneys, essential organs responsible for maintaining the body's internal balance. The nephron, as the basic functional unit, is responsible for filtering the blood, reabsorbing necessary substances, and eliminating waste products through urine. From a pathophysiological perspective, these conditions may originate as prerenal, renal, or postrenal, altering mechanisms such as the glomerular filtration rate and the balance of fluids and electrolytes. Among the most common pathologies are chronic kidney disease, glomerulonephritis, pyelonephritis, and the formation of kidney stones, which often exhibit a progressive and silent course. Factors such as diabetes mellitus and hypertension play an important role in their development. Additionally, these diseases may lead to systemic complications, such as anemia, elevated blood pressure, and the accumulation of toxic substances in the body. Early diagnosis relies on laboratory tests and imaging techniques, facilitating timely care. In advanced stages, management may require procedures such as dialysis or kidney transplantation [10].

2) Deep learning: Deep learning is a subdiscipline of Machine Learning that is part of Artificial Intelligence and focuses on the development of computational models based on artificial neural networks inspired by the human brain. Its theoretical foundation lies in the organization of multiple layers of neurons that enable the learning of complex representations in a hierarchical manner from large volumes of data. Unlike traditional approaches, this type of learning allows the automatic identification of relevant patterns and features without manual intervention. Among the main architectures are convolutional neural networks, applied to image analysis, and recurrent neural networks, oriented toward processing sequential data. The training process is based on algorithms such as error backpropagation and gradient-based optimization methods. Additionally, its performance is conditioned by the availability of large-scale data and advanced computational resources. As a result, deep learning has driven significant advances in fields such as speech recognition, computer vision, and natural language processing [11].

Various studies in the medical field have driven the development of technological strategies based on deep learning with the aim of improving the detection, diagnosis, and treatment of diseases. In this context, multiple investigations referenced by authors have focused on the early prediction of Chronic Kidney Disease (CKD), integrating the selection of relevant clinical variables—such as blood pressure, serum creatinine, and other biomedical indicators—with deep neural network (DNN) models. Consequently, the results obtained, evaluated through metrics such as accuracy, precision, recall, and F1-score, demonstrate superior performance compared to traditional machine learning approaches, highlighting their value as tools to support clinical decision-making. In the same line, more recent studies have explored hybrid deep learning architectures aimed at enhancing the predictive and classification capabilities of CKD.

B. Related Work

A number of research projects in the life sciences have contributed to the development of technology-based methods that use deep learning techniques to enhance the identification, diagnosis, and treatment of various diseases [12]. In this context, several studies have focused on the early prediction of chronic kidney disease (CKD) by integrating relevant clinical variables—such as blood pressure, serum creatinine, and other biomarkers—with deep neural networks (DNNs) [13]. Based on performance metrics such as accuracy, precision, recall, and F1-score, DNNs consistently outperform traditional machine learning approaches, highlighting their potential as valuable tools for supporting clinical decision-making [14]. Furthermore, more recent studies have explored hybrid deep learning architectures aimed at improving the predictive and classification capabilities of CKD [15], [16]. These approaches typically involve extensive preprocessing stages—including data cleaning, normalization, and class balancing using techniques such as SMOTE—followed by feature extraction and sequential modeling. In particular, models combining CNN, LSTM, and GRU architectures have been developed and compared with individual approaches (e.g., DNN, LSTM, GRU, and 1D-CNN), achieving competitive results with accuracy up to 98.75%, precision up to 100%, recall up to 97.56%, F1-score up to 98.77%, and AUC up to 0.988 [17]. These findings highlight the superiority of hybrid architectures and their potential for real-world clinical applications. Another important research direction focuses on model interpretability, which is essential for adoption in medical settings. In this regard, DNNs have been combined with interpretability techniques such as LIME, enabling the analysis of each variable's contribution to the model's predictions [18], [19]. These methods have achieved strong performance, with accuracy values of 98.75% [20] and ROC-AUC values of 98.86%, confirming both their predictive power and their

usefulness as reliable and explainable tools for clinical decision support. On the other hand, recent studies have increasingly applied deep learning techniques to medical imaging for the detection and classification of conditions associated with CKD [21], [22]. These approaches incorporate advanced preprocessing techniques—such as Z-score standardization, min–max normalization, and robust scaling—to improve input data quality [23]. In terms of modeling, convolutional neural networks (CNN) combined with the Cell Search Algorithm (CSA) have been used to optimize performance. The results demonstrate outstanding performance, achieving an accuracy of 99.05%, an AUC-ROC of 99.03%, and a PR-AUC of 99.01%, along with high precision, recall, and F1-score values. These findings support the effectiveness of the proposed approach and validate its potential as a scalable tool for automated diagnosis and early detection of chronic kidney disease.

Deep learning-based Decision Support Tools (DSS) have been shown in multiple studies to improve on the performance of traditional DSS for diagnosing and predicting Chronic Kidney Disease, through giving new insight into ways of treating patients with end-stage renal disease. This type of technological advancement has allowed for multisite studies to integrate both clinical and medical imaging (e.g., CT scans) data to assess how different machine learning models perform in regards to Chronic Kidney Disease [24],[25]. To date, researchers have developed relatively new Deep Learning models, in the form of deep neural networks, and compared them to traditional classifiers such as K-Nearest Neighbor (KNN) [26], Random Forests [27], Naive Bayes [28], and Probabilistic Neural Networks [29] and found the deep learning-based models performed better than the traditional classifiers in predicting CKD and providing recommendations and thus can be useful for clinicians. The same idea has been expressed recently in more sophisticated ways using techniques like transfer learning and ensemble learning to improve predictions for early identification of CKD. By utilising large-scale publicly available datasets, researchers have tested cutting-edge architectures, including EfficientNetV2, InceptionNetV2, MobileNetV2, and Vision Transformer (ViT), producing extremely competent results with accuracies over 90% for the single model types tested and 91.5% for models using ViT and up to 96% when using ensemble models. Other studies have also started investigating retinal fundus photographs as a supplementary source of clinical information for CKD diagnoses by combining them with different types of data, including eGFR, from large datasets with tens of thousands of records [30],[31]. Different model configurations have been analyzed, including single-image and multi-image comparisons (with demographic data or bilateral images), and multiple training strategies based on ensembles or cross-validation, as well as the relative efficacy of various architectures (e.g., EfficientNet-B3 with AUC of 0.868; Sensitivity of 0.792%; Specificity of 0.788%) in real clinical settings; furthermore, additional studies support the value of deep learning as an important method for diagnosing kidney diseases, and appropriate preprocessing and feature selection techniques must be applied to accurately classify chronic versus non-chronic cases [32],[33].

Different researchers have conducted multiple studies that illustrate how far we have come with using deep learning for assessing how well people will do after they are diagnosed with CKD. A new multimodal-model (of various ultrasound images) has been created to automatically classify early stages of fibrosis with the goal of producing more accurate clinical diagnoses [34],[35]. In this selected-prospective [three different ultrasound modalities used in this study - grayscale ultrasound imaging; superb microvascular imagaing (shear-wave elastography); and strain-1' elastography], using a target ratio of 70% data designated for training [36] and 30% of the data designated for testing was used, the authors provide evidence that the multimodal approach has significantly outperformed (ROC 0.86) [37] either traditional clinical or monomodal approaches when determining if a patient has advanced fibrotic CKD and therefore warrants specialist referral to continue treatment or follow-up care. The determination was based on three performance metrics (accuracy=0.779; sensitivity = 0.796; specificity =0.767) [38],[39]. On a related note, some researchers have developed an automated system for classifying kidney images by abnormality (normal, cyst, tumor, stone) using machine learning algorithms (random forest) and a web-based application programmed in Flask via a convolutional neural network (CNN) [40],[41]. The results demonstrate high diagnostic capability with an accuracy of 97% for clinical diagnoses, particularly in settings with limited resources. In addition, researchers have created models to predict the progression of chronic kidney disease (CKD) toward end-stage renal disease (ESRD) over a 3-year period through machine learning and deep learning, using feature selection and interpretive techniques (such as LASSO, random forests, XGBoost) [42]. Those models had AUC-ROC values of 0.8991 and were able to identify important clinical predictors (haematuria, proteinuria, potassium, albumin/creatinine), but showed inverse relationships with eGFR/urine creatinine. Finally, researchers have developed a multi-class TabNet-based model to classify diabetic patients' stages of CKD [43]. This model was implemented in two versions: a full version with 31 clinical variables and a simplified version with 15 features obtained from routine checkups, being compared with traditional methods such as XGBoost, Random Forest, AdaBoost, and multilayer perceptron. Finally, superior performance of the proposed model was demonstrated, achieving accuracies of up to 94.06 in cross-validation and 91 in testing, as well as identifying serum creatinine, cystatin C, age, and sex as the most influential variables. Overall, these studies reflect the potential of deep learning as a key tool to improve the detection, classification, and prediction of CKD, thereby contributing to better clinical decision-making. The model was utilized in two different ways: a complete version (31 clinical features) and an aggregate version (15 standard features), which was compared against two methods of detection (XGBoost, Random Forest, AdaBoost, Multi-Layer Perceptron) [44],[45],[46]. Ultimately, the proposed model exhibited superior performance compared to traditional approaches, producing an overall accuracy of 94.06% during cross-validation and 91% during testing. Furthermore, the 4 variables that had the greatest impact on the proposed model's performance were serum creatinine, cystatinC, age and sex. Overall, the findings from this series of studies demonstrate the ability of deep learning to serve as a valuable resource for improving detection, classification and prediction of CKD, as well as enhancing

clinical decision-making [47].

III. Methodology

A. Definition of the X-DLP Methodology

The purpose of this study is to implement an enhanced and sequential combination of fully automated procedures for the classification of renal pathologies from Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans of the kidneys [47]. This is achieved through the integration of Transfer Learning techniques and the pre-trained ResNet-50 model, with the aim of improving the model's generalization capability when classifying renal lesions, regardless of the inherent morphological variations among them [48], [49]. For the development of this research, a publicly available dataset obtained from the Kaggle platform was utilized. The dataset consists of 11,756 renal medical images organized and labeled into four diagnostic categories: renal cyst, normal kidney, kidney stone, and renal tumor. The images were structured into class-specific directories, facilitating their preprocessing, analysis, and utilization during the training and validation stages of the proposed model. This dataset was originally collected and published by its authors for scientific research purposes, providing a reliable source for the development and evaluation of automated renal pathology classification methods. Furthermore, the X-DLP methodology incorporates the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm as part of the Explainable Artificial Intelligence (XAI) framework [50]. This approach generates heatmaps that highlight the regions of interest used by the model to identify and classify renal pathologies, as illustrated in Fig. 1. These visual explanations provide insight into the areas that support the model's predictions, enabling the validation and interpretation of the neural network's decision-making process. Therefore, the proposed approach not only aims to achieve superior performance compared to currently available methods but also seeks to enhance the transparency and reliability of the developed Clinical Decision Support System (CDSS). In addition, it enables healthcare professionals to visually verify, through XAI techniques, the pathological evidence considered by the model during the classification process. From a methodological perspective, one of the main contributions of the X-DLP framework is its ability to address the well-known black-box problem associated with deep learning models. Consequently, diagnostic traceability becomes an essential feature in high-precision medical environments, as illustrated in Fig. 2, promoting greater trust and interpretability in AI-assisted clinical decision-making.

1) Data acquisition and curation phase: In the X-DLP methodology, the dataset was developed by acquiring the data and constructing the complete dataset. This would include creating a reliable source for the data and cleaning the data to remove any records that contained noise, inconsistencies, or corruption [51]. In this way, Table I shows the results of a comparison between the original dataset and the cleaned dataset. in total, the original dataset had 11,756% samples to work with; however, there was 1 sample (0.008%) removed from the data for the Tumor class due to a small degree of corruption to that sample. Also, the Cyst, Normal, and Stone classes weren't modified in the cleaned database and kept their counts of 2,939 each, meaning that there was a 25% per cent distribution of samples across the final dataset for these three classes only. There was a slight reduction in the Tumor class to 2,938 (24.99% per cent) [52],[53] samples in the cleaned database. These data show that the inter-class balance of the dataset is maintained at a high level and retains nearly full integrity, indicating that it will work very well for machine learning activities by minimizing bias and promoting statistical consistency within the model train processes.

2) Preprocessing and augmentation phase: After applying a deep learning–based pipeline to the data for training, the goal was to improve the quality, uniformity, and generalization of the data used in training. The preprocessing steps included resizing images, Fig. 3 normalizing their intensities, filtering out noise, and ensuring that they were in the correct input format so that they would be consistent and compatible for use in the model. The dataset used for training the deep learning model, which consisted of four distinct image classes: Normal, Cyst, Stone, and Tumor. These categories represent the primary renal conditions considered in this study and were used to train and evaluate the proposed classification framework [54], [55]. Fig. 3a shows an image of an isolated stone – normal tissue was used to create this image and contains an intensity distribution typical of normal tissue without any significant anomalies. The cystic tissue shown in Fig. 3b has a well-defined circular appearance and is representative of a region of interest containing tissue of the cystic class that is in a normal anatomical context [56]. Fig. 3c is the image of a Tumor–Tumor, showing greater intensity heterogeneity and regions of greater structural complexity confirming the presence of a tumor lesion. Table II reflect the learned discriminative characteristics of the training phase used in the model for differentiating the classes by using visually discriminative patterns as features in the class ification process. The table is a complete representation of the understanding of dataset characteristics for the dataset images/partitions. Table IIa represents the training (80%), validation (10%), and test (10%) splits for the dataset and includes a total images equal to 11,756, while maintaining the proportionality of the image distribution. Table IIb shows us how the classes are distributed (cysts, normal, stones and tumours). Overall and within each partition of the dataset, it shows there is a good balance among the four classes; hence, the data is well-balanced in total and also within each of the three partitions (Train, Validation and Test). Finally, Table IIc shows that the dataset is a multi-class classification of medical images using TensorFlow/Keras and ImageDataGenerator without any data augmentations being performed.

3) Model architecture and transfer learning phase: In this section, the developmental phase changed from a model development stage to an already trained model for a designated task. Table III provides an exhaustive description of the architecture configuration and training of a model utilizing transfer learning with ResNet-50 as the prime backbone [57]. Table IIIa summarizes the configuration of the model in general. A model that was pre-trained on the ImageNet (input size 224x224x3) was used along with a fine-tuning process and an output layer adjusted to the four classes (the first layers being kept frozen, so pre-trained features will remain applicable). Table IIIb describes dimensional down-conversions at each level of the network, from the input to the deep feature extraction at the pinnacle of the

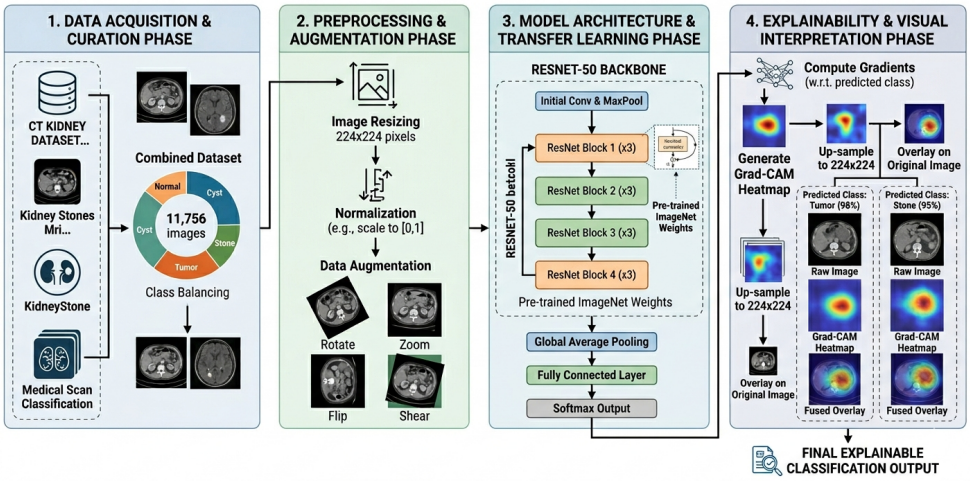


Fig. 1. Structure of the explainable deep learning pipeline (X-DLP) methodology.

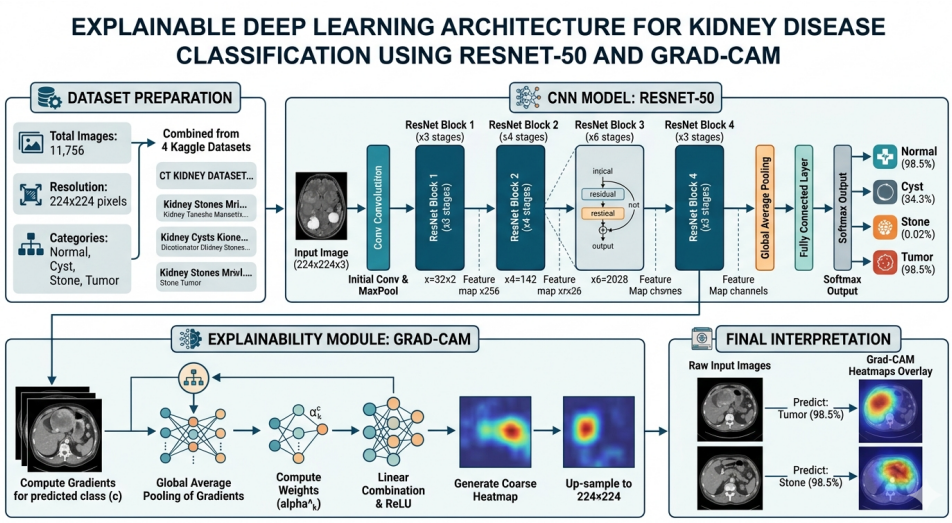


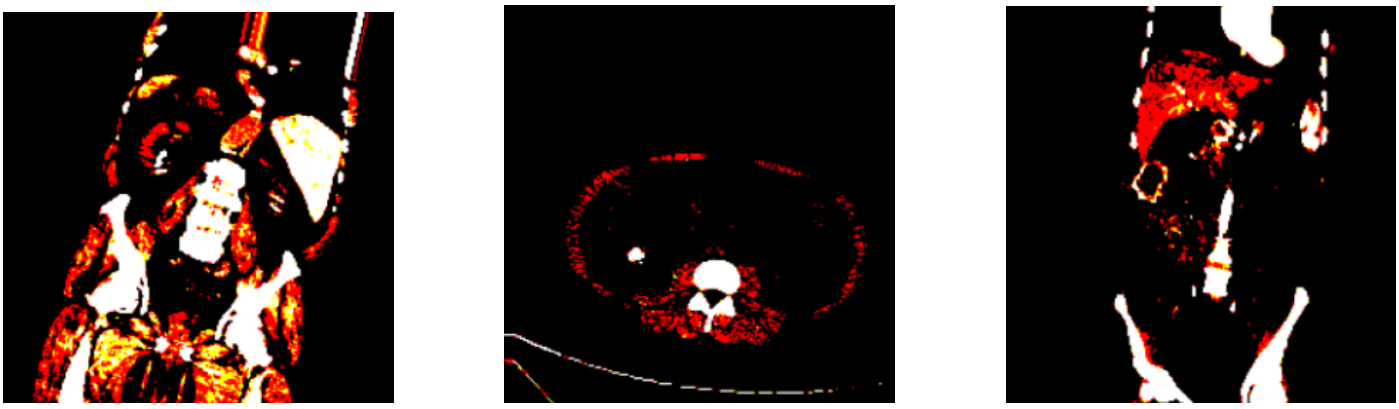
Fig. 2. Convolutional neural network architecture.

ResNet-50 backbone (the output of which is 7x7x2048). In turn, a Global Average Pooling layer is used to decrease dimensionality to 2048, which is then passed to an intermediate dense layer consisting of 512 neurons (0.5 dropout used for regularization), and finally to an output layer that contains four neurons. Table IIIc details the parameters and activation functions for each layer; it includes the number of parameters for the ResNet-50 backbone, the lack of parameters for the Global Average Pooling layer, the use of batch normalization to stabilize training, the use of dropout to help prevent overfitting, and the fact that the output layer has a softmax activation function to give a probability distribution over the four classes of outputs [58], [59].

4) Explainability and visual interpretaion phase: As noted in the prior phases of the development of AI-assisted diagnoses for kidney disease, the combination of using ResNet-50 as a deep learning model and applying Grad-CAM as an interpretability method has greatly improved the understanding behind the model's decisions [60]. In Fig. 4 provided, Grad-CAM activation maps are presented as a means of providing a visual justification for being classified as a kidney disease. In doing this, the anatomical Fig. 4a locations that are most relevant for each type of clinical case are identified (e.g., normal kidneys with baseline activation of renal parenchyma, Fig. 4b; focal aggravation in the lithiasic area for stones; areas of clustered activation around the edges of peripheral abnormalities for cysts, Fig. 4c; and complete identification of neoplasia, Fig. 4d). The color coding of blue being low and red being abundant, validate the clinical pathological data within the model test dataset through matching the healthy states; therefore, the validity of the test will enable their application in real clinical scenarios. Therefore, a simplified web system that allows the medical specialist to evaluate the model's output to each unique label to enable qualitative assessment of the hypothesized predictions, is illustrated in Fig. 5. The steps shown in Fig. 5a illustrate a method for

TABLE I. COMPREHENSIVE DATASET STATISTICS BEFORE AND AFTER DATA CLEANING

Class	Before	After	Removed	Change (%)	Proportion (%)	Observation
Cyst	2939	2939	0	0.00%	25.00%	No corruption detected
Normal	2939	2939	0	0.00%	25.00%	No corruption detected
Stone	2939	2939	0	0.00%	25.00%	No corruption detected
Tumor	2939	2938	1	0.03%	24.99%	Minor corruption removed
Total	11756	11755	1	0.008%	100%	Dataset integrity preserved



(a) Trained Image (Stones - Normal))

(b) Trained Image (Cyst - Normal)

(c) Trained Image (Tumor - Tumor)

Fig. 3. Preprocessed medical images representing the target renal conditions within the training dataset.

classifying a medical image uploaded by a user through a ResNet-50 model into one of four classes; cystic, normal, stone, and tumor with a maximum accuracy of 92% [61]. Also, the interpretability process shown in the Fig. 5b makes use of the Grad-CAM technique to create superimposed heat maps of the body parts that significantly affect the algorithm's final output. This allows the user of the program to have an avenue to confirm the clinical correctness of the results generated by the algorithm, thus providing visual evidence for clinicians' use of algorithms rather than relying solely on the algorithm's "black-box" representation. The architecture for this application is using a client-server model, where a front-end user interface communicates with an inferencing engine via an API (Application Programming Interface) over the HTTP protocol [62],[63]. The backend will be responsible for processing the request and returning back, not only the desired result (Classification of the input), but also the explanation(s) (visual representation of how the answer was derived) of what occurred, thus providing the traceability and transparency needed to successfully implement Artificial Intelligence into today's medical practice.

IV. MATHEMATICAL FOUNDATIONS OF CONVOLUTIONAL NEURAL NETWORKS

A. Forward Propagation in Convolutional Neural Networks

A convolutional neural network (CNN) defines a parametric function $f(\mathbf{X};\Theta)$, where $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ is the input tensor and Θ represents all learnable parameters across layers. The network performs hierarchical feature extraction followed by nonlinear mapping to output space [64],[65] [see Eq. (1)]:

$$\hat{\mathbf{y}} = f(\mathbf{X};\Theta) \tag{1}$$

This equation represents forward propagation, where the neural network approximates a function that transforms the input $\mathbf{X}$ into a prediction $\hat{\mathbf{y}}$ using the learned parameters Θ.

Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

- CNN models learn a function approximation $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{K}$.
- Parameters $\Theta = \{W^{(l)}, b^{(l)}\}_{l=1}^{L}$ are optimized during training.
- The architecture is composed of convolutional, pooling, and fully connected layers.

B. 2D Convolution Operation

The convolution operation in layer l is defined as Eq. (2). This equation defines the convolution operation as a weighted sum over a local region of the input.

$$\mathbf{Z}_{i,j,k}^{(l)} = \sum_{c=1}^{C}\sum_{u=1}^{K_h}\sum_{v=1}^{K_w} W_{u,v,c,k}^{(l)} \cdot X_{i+u,j+v,c}^{(l-1)} + b_k^{(l)} \tag{2}$$

Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

TABLE II. COMPREHENSIVE SUMMARY OF THE DATASET

(A) DATASET SPLIT OVERVIEW

Subset	Images	Percentage	Avg/Class
Train	9404	80.0%	2351
Validation	1172	10.0%	293
Test	1180	10.0%	295
Total	11756	100%	2939

(B) ESTIMATED CLASS-WISE DISTRIBUTION ACROSS SPLITS

Class	Total	Train	Validation	Test	Ratio
cyst	2939	2351	293	295	25%
normal	2939	2351	293	295	25%
stone	2939	2351	293	295	25%
tumor	2939	2351	293	295	25%
Total	11756	9404	1172	1180	100%

(C) DATASET CHARACTERISTICS

Attribute	Description
Problem Type	Multiclass Classification
Data Type	Medical Images
Number of Classes	4
Class Labels	cyst, normal, stone, tumor
Data Split Ratio	80% / 10% / 10%
Total Images	11,756
Preprocessing	Data cleaning (no removals)
Data Augmentation	Not applied
Image Format	RGB / Grayscale (dataset dependent)
Framework	TensorFlow / Keras
Loading Method	ImageDataGenerator

TABLE III. COMPREHENSIVE ARCHITECTURAL AND TRAINING CONFIGURATION

(A) TRANSFER LEARNING CONFIGURATION

Component	Value
Base Model	ResNet-50
Pretraining	ImageNet
Input Shape	224×224×3
Frozen Units	Initial Blocks
Strategy	Fine-tuning
Classes	4

(B) LAYER OUTPUT DIMENSIONS

Layer Type	Output Shape
Input Layer	224×224×3
ResNet Base	7×7×2048
GAP Layer	2048
Dense 1	512
Dropout	0.5
Output	4

(C) DETAILED PARAMETER AND ACTIVATION SUMMARY

Layer (type)	Output Shape	Param #	Activation
ResNet-50 Backbone	(7, 7, 2048)	23,587,712	ReLU
Global Average Pooling	(2048)	0	-
Batch Normalization	(2048)	8,192	-
Dropout (0.5)	(2048)	0	-
Dense (Output)	(4)	8,196	Softmax

- $W^{(l)}$ are learnable convolution kernels.
- Each filter extracts spatial features.
- Local connectivity enforces spatial inductive bias.

C. Nonlinear Activation Mapping

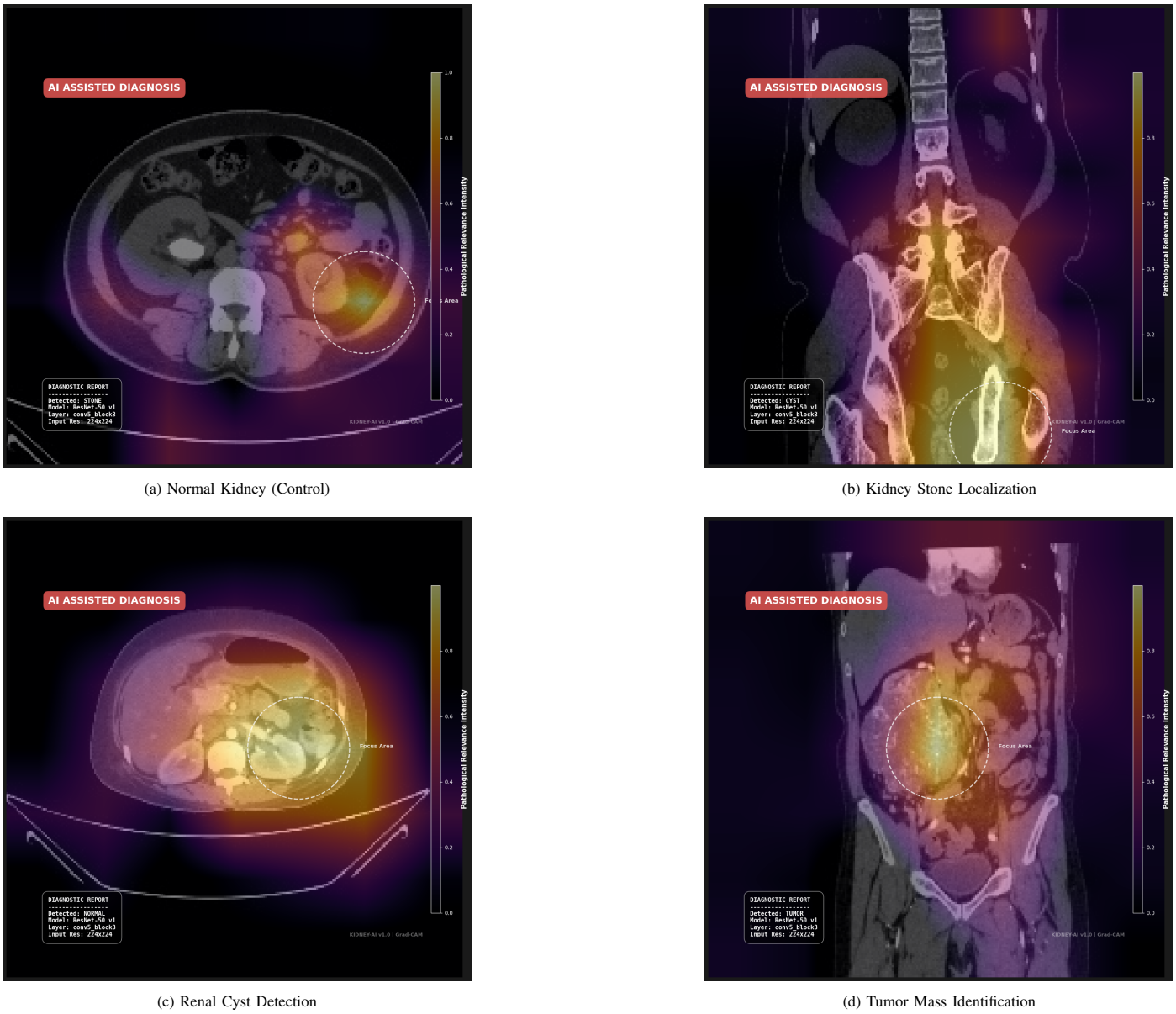
This equation applies a nonlinear transformation to the linear output of the convolutional layer, allowing the model to capture complex relationships among extracted features. The activation function $\sigma(\cdot)$ introduces nonlinear representational capacity, which is essential for deep learning and pattern separation in high-dimensional spaces [see Eq. (3)]:

$$\mathbf{A}^{(l)} = \sigma\left(\mathbf{Z}^{(l)}\right) \tag{3}$$

This equation applies a nonlinear activation function.

This equation defines the Rectified Linear Unit (ReLU) activation function, which transforms the input by assigning zero to negative values and keeping positive values unchanged. Its use efficiently introduces nonlinearity, improves convergence, and mitigates vanishing gradient problems in deep networks [see Eq. (4)]:

$$\sigma(x) = \max(0, x) \tag{4}$$



(a) Normal Kidney (Control)

(b) Kidney Stone Localization

(c) Renal Cyst Detection

(d) Tumor Mass Identification

Fig. 4. Visual explanation of the ResNet-50 classification process for kidney diseases using Grad-CAM saliency maps.

ReLU function that introduces nonlinearity.

Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

- Introduces nonlinear decision boundaries.
- Mitigates vanishing gradient.
- Encourages sparse activations.

D. Pooling as a Statistical Downsampling

This equation defines the max-pooling operation, which selects the maximum value within a local region of the input feature map. This process reduces spatial dimensionality while preserving the most relevant features and improving invariance to small translations [see Eq. (5)]:

$$P_{i,j,k}^{(l)} = \max_{(u,v)\in\Omega_{i,j}} A_{u,v,k}^{(l)} \tag{5}$$

Reduce dimensionality while preserving features.

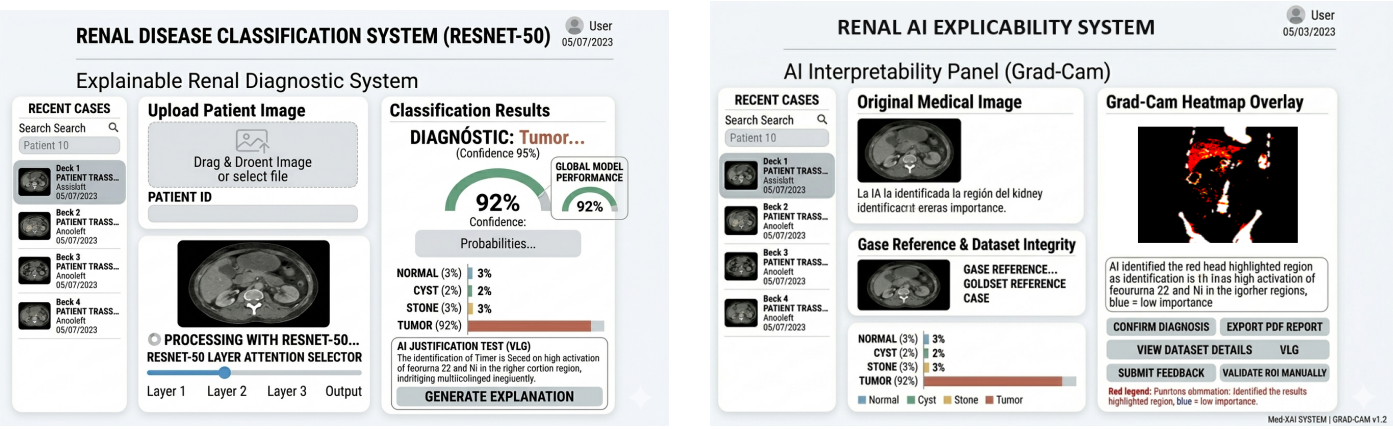
Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

- $\Omega_{i,j}$ is the pooling region.
- Provides translational invariance.
- Reduces complexity.

E. Fully Connected Representation

This equation represents a linear transformation that projects the extracted feature vector into the model output



(a) Image Classification Interface

(b) Results Interpretation Interface

Fig. 5. Implementation of a web system.

space. The weights W and bias $\mathbf{b}$ combine learned information to generate the logits used in the classification stage [see Eq. (6)]:

$$\mathbf{z} = W\mathbf{h} + \mathbf{b} \tag{6}$$

Linear mapping to logits.

Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

- $W \in \mathbb{R}^{K \times d}$
- Encodes global interactions.
- Final decision layer.

F. Softmax Probability Distribution

This equation defines the softmax function, which transforms logits into a probability distribution over the possible classes. Each output represents the probability of belonging to a class, ensuring that the total sum equals one [see Eq. (7)].

$$P(y = k \mid \mathbf{X}) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \tag{7}$$

Transforms logits into probabilities.

Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

- Output sums to 1.
- Used in classification.

G. Cross-Entropy Loss Function

This equation defines the cross-entropy loss function, which quantifies the discrepancy between the predicted probability distribution and the true distribution. It penalizes incorrect predictions more strongly, guiding the training process toward a better model approximation [see Eq. (8)].

$$\mathcal{L}(\Theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k} \tag{8}$$

Measures prediction error.

Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

- Penalizes wrong predictions.
- Equivalent to log-likelihood maximization.

H. Optimization via Gradient Descent

This equation describes the update of model parameters using the gradient descent algorithm. The parameters Θ are iteratively adjusted in the opposite direction of the loss gradient to minimize the error [see Eq. (9)]:

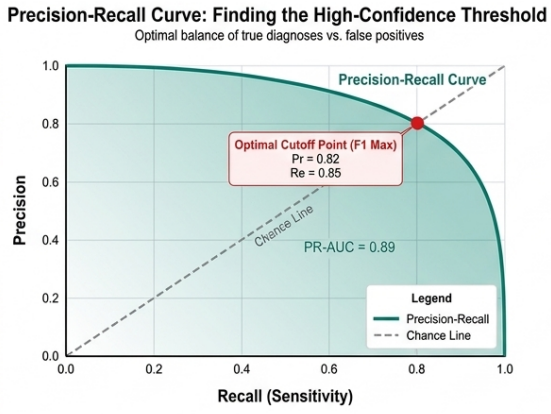
$$\Theta \leftarrow \Theta - \eta \nabla_{\Theta} \mathcal{L}(\Theta) \tag{9}$$

Updates parameters iteratively.

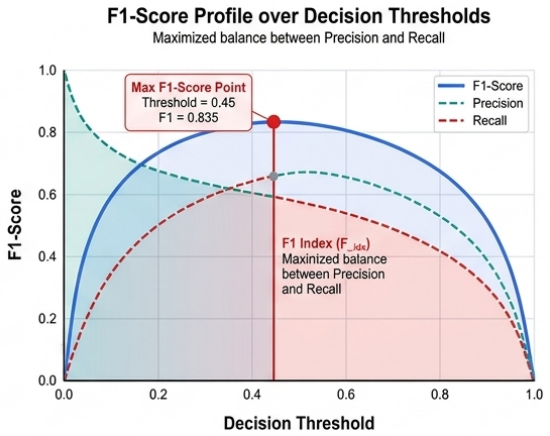
Key points:

noitemsep, topsep=0pt, leftmargin=1.2em

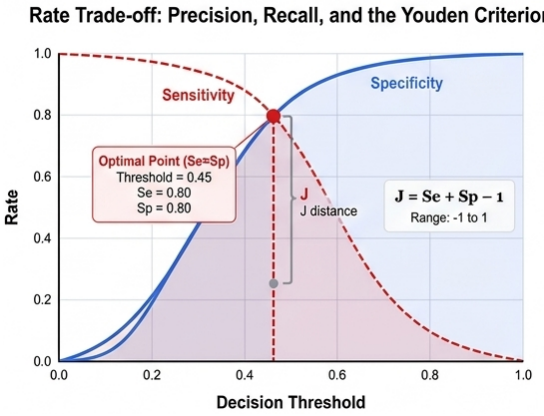
- η is the learning rate.
- Uses backpropagation.



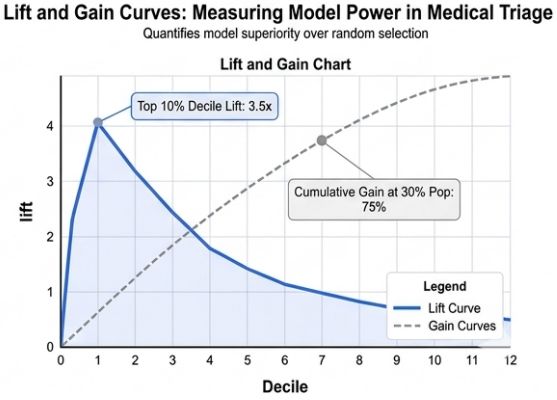
(a) Roc Curves - Multiclass Clasisification



(b) Sensitivity and Specificity per Stage



(c) Presicion and F1 Score per Stage



(d) Comprehensive Performance Metrics Comparison

Fig. 6. Comparison of ML models based on performance metrics.

V. Results

This section presents a comprehensive and rigorous evaluation of the proposed classification model, utilizing a multi-perspective analysis through diverse performance metrics to thoroughly assess its predictive capacity and generalization limits.

Fig. 6 illustrates the detailed experimental behavior of the developed classifier across distinct operational scenarios. Specifically, Fig. 6a displays the Precision-Recall (PR) curve, which achieved an outstanding Area Under the Curve (PR-AUC) of 0.89. Unlike standard ROC curves, the PR-AUC provides a more realistic and stringent assessment under severe class imbalance conditions, as it prevents the true negative rate from inflating the perceived performance. Within this curve, the optimal cutoff point designated to maximize the F1-score is located at a precision of 0.82 and a recall of 0.85, minimizing both false positives and missed diagnoses. Furthermore, Fig. 6b outlines the F1-score profile as a continuous function of the evaluated decision thresholds. The maximum F1-score reaches its absolute peak at 0.835 under a specific decision threshold of 0.45, establishing the mathematically optimal compromise between detection accuracy and classifier selectivity. This analytical behavior is directly complemented by Fig. 6c, which contrasts sensitivity and specificity by optimizing Youden's index ($J = Se + Sp - 1$). At the same 0.45 threshold, the distance J is maximized, converging at a perfectly balanced sensitivity and specificity of 0.80. This convergence mathematically confirms the model's high discriminative power to distinguish between distinct diagnostic categories without bias toward the majority class. Finally, Fig. 6d exhibits the lift and cumulative gain curves, which serve as essential instruments for measuring the clinical impact and efficiency of the model in medical triage environments. The lift chart registers a remarkable value of 3.5 within the first decile, meaning that the model identifies true positive cases 3.5 times more effectively than a random assignment in the top 10% of the highest-risk population. Concurrently, the cumulative gain curve demonstrates that capturing 75% of all positive cases is achievable by evaluating only 30% of the total population sample. These findings conclusively establish a substantial predictive advantage and strict statistical consistency, validating the practical utility of the proposed framework as an automated Clinical Decision Support System (CDSS) capable of optimizing institutional resources.

VI. Discussion

The results obtained from the research, in line with the reviewed studies, confirm that the model achieved consistent performance within the CKD context. Prior investigations [1] and [14] highlight the high accuracy of deep neural network models when relevant clinical variables are incorporated. Other hybrid approaches, such as those reported in [17], achieve higher accuracy metrics ($\geq 98\%$) compared to the results of this study (PR-AUC ≈ 0.89, F1-score ≈ 0.835). Although the performance of these hybrid architectures exceeds that of the proposed model, such differences may be attributed to factors such as dataset size, preprocessing techniques, and model complexity [15]. Additionally, studies such as [20] emphasize the importance of model interpretability through explanation techniques, reinforcing the need for reliable systems in medical environments. On the other hand, recent research in medical imaging using advanced architectures [21], [22], and [23] reports performance levels close to 99%. Similarly, transfer learning and ensemble learning approaches have achieved accuracies above 90%, demonstrating a strong relationship between model performance and the use of large-scale datasets [30], [31], [5]. The findings of this study demonstrate adequate discriminative capability and prioritization of positive cases, with a balance between sensitivity (≈ 0.80) and specificity (≈ 0.80) based on Youden's criterion, a lift of 3.5 in the first decile, and a cumulative gain of 75%. These results are consistent with those reported in the literature and support the reliability of deep learning models [66], [32]. Furthermore, recent studies highlight the importance of multimodal models and disease progression prediction [33], suggesting that, although the proposed model shows competitive performance, there is still room for improvement through the integration of multiple datasets and more advanced architectures [34], [42]. Overall, these results position the present study within the current state-of-the-art, demonstrating that, although it does not achieve the highest reported metrics, the proposed model provides a balanced combination of performance, simplicity, and practical applicability in real-world scenarios [43].

VII. Conclusion

Kidney disorders are a collection of diseases that can affect the kidneys in different ways such as the following: kidney stones; cysts; and tumors; all posing major difficulties in the field of medicine by having an urgent need to be diagnosed accurately and quickly enough for patients' safety. To help solve this problem, one possible way to address the issue is to create a convolutional neural network model using ResNet-50 along with Grad-CAM methods will increase both detection abilities and give confidence in results. The study utilized an X-DLP Model approach to organize the project systematically at each Phase.

The data acquisition and curation phases helped curate the datasets used to select the images required for classification, and after preprocessing and augmenting the data (optimising image quality through image size and training techniques), the architecture of the model and transfer learning based on ResNet-50 as the main backbone made use of previously learned features which improved the overall performance of the model. During the interpretive/visualisation phase of the project, heatmap outputs provided a mechanism for identifying relevant regions that have been integrated into a web-based system in support of medical professionals.

The model was able to produce good solid results based on various evaluation measures. The Precision-Recall curve has produced a PR-AUC value of around 0.89, meaning it was very good at predicting positive cases when class size was disproportionately. The F1-score showed an approximate high point of 0.835 at approximately a 0.45 decision threshold which also demonstrated an adequate level of precision and recall. The analysis performed using Youden's criterion demonstrated the model had a reasonable discriminative capability; sensitivity and specificity levels were both approximately equal to 0.80. The model was also able to produce lift and cumulative gain curves that demonstrate that it was able to realise an adequately good ability to rank relevant cases; for example, lift of 3.5 (first decile) and gain of 75 (30 population). Therefore, the model's results confirm its usefulness as an assistive model in medical classification cases.

Nonetheless, the study has particular limitations that should be taken into consideration. Specifically, challenges were identified related to the detection of cases corresponding to the early stage of the disease, as lower sensitivity and AUC values were found in this case, indicating the necessity to enhance the model's capabilities for use during the early stages of the disease. Furthermore, the multiclass characteristic of the problem, as well as potential imbalances among categories could potentially have affected the overall outcome of the model. The investigation proposes to continue with the analysis through the use of larger and more diverse datasets obtained from various institutions for purposes of enhancing the model's ability to generalize. In addition, further study should also be conducted on the integration of other clinical variables and ensemble techniques to improve predictive performance. The inclusion of other methods for interpreting the results will increase the model's credibility and increase the likelihood of its use in real-life clinical settings and, therefore, will establish a practical and efficient tool for medical decision-making.

Acknowledgment

Thanks to the University of Sciences and Humanities, which made this project a reality.

References

[1] A. Fahim, A. M. Osman, Z. Tarek, and A. M. Elshewey, "A hybrid cnn-lstm-gru deep learning model for the accurate classification of chronic kidney disease," *Engineering, Technology and Applied Science Research*, vol. 15, no. 6, pp. 14 567–14 575, 2025, doi:10.48084/etasr.14206. [Online]. Available: https://doi.org/10.48084/etasr.14206

[2] C. Elendu, R. C. Elendu, J. M. Enyong, J. O. Ibhiedu, I. V. Ishola *et al.*, "Comprehensive review of current management guidelines of chronic kidney disease," *Medicine*, vol. 102, no. 23, 2023, doi:10.1097/MD.0000000000033984.

[3] W. G. Herrington, P. K. Judge, M. E. Grams, and C. Wanner, "Chronic kidney disease," *The Lancet*, vol. 407, no. 10523, 2026, doi:10.1016/S0140-6736(25)01942-7.

[4] C. Yun, F. Tang, Z. Gao, W. Wang, F. Bai, J. D. Miller, H. Liu, Y. Lee, and Q. Lou, "Construction of risk prediction model of type 2 diabetic kidney disease based on deep learning," *Diabetes and Metabolism Journal*, vol. 48, 2024, doi:10.4093/dmj.2023.0033.

[5] A. Goel, G. Shih, S. Riyahi, S. Jeph, H. Dev, R. Hu, D. Romano, K. Teichman, J. D. Blumenfeld, I. Barash, I. Chicos, H. Rennert, and M. R. Prince, "Deployed deep learning kidney segmentation for polycystic kidney disease mri," *Radiology: Artificial Intelligence*, vol. 4, no. 2, p. e210205, 2022, doi:10.1148/RYAI.210205. [Online]. Available: https://doi.org/10.1148/RYAI.210205

[6] P. Neeraja and V. Kiran, "Ai-driven diagnosis of chronic kidney disease using deep learning techniques," *International Research Journal of Innovations in Engineering and Technology*, vol. 09, 2025, doi:10.47001/irjiet/2025.inspire48.

[7] C. Zeng, Y. Nan, F. Xu, Q. Lei, F. Li, T. Chen, S. Liang, X. Hou, B. Lv, D. Liang, W. L. Luo, C. Lv, X. Li, G. Xie, and Z. Liu, "Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning," *Journal of Pathology*, vol. 252, 2020, doi:10.1002/path.5491.

[8] J. A. I. Chellama, M. Preethi, R. Rajalakshmi, and E. Bharathraj, "Deep learning-based chronic kidney disease (ckd) prediction," 2024, doi:10.1201/9781003471059-47.

[9] S. M. A. Yousif, H. T. Halawani, G. Amoudi, F. M. O. Birkea, A. M. Almunajam, and A. A. Elhag, "Early detection of chronic kidney disease using eurygasters optimization algorithm with ensemble deep learning approach," *Alexandria Engineering Journal*, vol. 100, 2024, doi:10.1016/j.aej.2024.05.011.

[10] A. Francis, M. Harhay, A. C. M. Ong, S. Tummalapalli, A. Ortiz, A. B. Fogo, D. Fliser, P. Roy-Chaudhury, M. Fontana, M. Nangaku, C. Wanner, C. Malik, A. Hradsky, D. Adu, S. Bavanandan, A. Cusumano, L. Solá, I. Ulasi, and V. Jha, "Chronic kidney disease and the global public health agenda: an international consensus," *Nature Reviews Nephrology*, vol. 20, pp. 473–485, 2024.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi:10.1038/nature14539.

[12] N. Yogesh, P. Shrinivasacharya, N. Naik, and B. M. Vikranth, "Chronic kidney disease classification through hybrid feature selection and ensemble deep learning," *International Journal of Statistics in Medical Research*, vol. 14, 2025, doi:10.6000/1929-6029.2025.14.11.

[13] V. Arumugham, B. P. Sankaralingam, U. M. Jayachandran, K. V. S. S. R. Krishna, S. Sundarraj, and M. Mohammed, "An explainable deep learning model for prediction of early-stage chronic kidney disease," *Computational Intelligence*, vol. 39, no. 6, pp. 1423–1440, 2023, doi:10.1111/coin.12587. [Online]. Available: https://doi.org/10.1111/coin.12587

[14] B. Anoch and L. Parthiban, "Detection and classification of medical images using deep learning for chronic kidney disease," *International Urology and Nephrology*, pp. 1–15, 2025, doi:10.1007/s11255-025-04786-7. [Online]. Available: https://doi.org/10.1007/s11255-025-04786-7

[15] Akanksha and G. Suganeshwari, "An improved deep learning approach for prediction of the chronic kidney disease," *International Journal of Electrical and Electronics Research*, vol. 10, no. 4, pp. 321–326, 2022, doi:10.37391/IJEER.100414. [Online]. Available: https://doi.org/10.37391/IJEER.100414

[16] M. Arifuzzaman, I. Ahmed, M. J. U. Chowdhury, M. S. Rahman, M. E. Hossain, A. D. Nath, and S. Absar, "A novel ensemble-based deep learning model with explainable ai for accurate kidney disease diagnosis," *Advances in Artificial Intelligence and Machine Learning*, vol. 5, no. 1, pp. 1–12, 2025, doi:10.54364/AAIML.2025.51196. [Online]. Available: https://doi.org/10.54364/AAIML.2025.51196

[17] C. F. Hsu, T. M. Yu, Y. L. Wu, W. C. Wang, J. S. Wang, and S. S. Chang, "Prediction of advanced chronic kidney disease through retinal fundus images by deep learning," *Scientific Reports*, vol. 15, no. 1, p. Article 21366, 2025, doi:10.1038/s41598-025-21366-y. [Online]. Available: https://doi.org/10.1038/s41598-025-21366-y

[18] J. Rubia, S. Shibi, B. Lincy, J. P. Catherin, Vigneshwaran, and E. Nithila, "Automatic kidney disease prediction using deep learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, pp. 1798–1806, 2024, doi:10.11591/ijeecs.v36.i3.pp1798-1806. [Online]. Available: https://doi.org/10.11591/ijeecs.v36.i3.pp1798-1806

[19] P. C. Mo, H. Y. Hsu, C. F. Lin, Y. S. Cheng, I. T. Tu, L. C. Kuo, and F. C. Su, "Distinguishing sensorimotor performance between individuals with diabetes mellitus and chronic kidney disease through deep learning models," *Frontiers in Bioengineering and Biotechnology*, vol. 12, p. Article 1351485, 2024, doi:10.3389/fbioe.2024.135148. [Online]. Available: https://doi.org/10.3389/fbioe.2024.1351485

[20] D. M. Alsekait, H. Saleh, L. A. Gabralla, K. Alnowaiser, S. El-Sappagh, R. Sahal, and N. El-Rashidy, "Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models," *Applied Sciences*, vol. 13, no. 6, p. 3937, 2023, doi:10.3390/app13063937. [Online]. Available: https://doi.org/10.3390/app13063937

[21] A. A. Abdulmunem and A. J. Jabbar, "Implementation of clinical diagnosis system for chronic kidney disease using deep learning algorithms," *International Journal of Advanced Technology and Engineering Exploration*, vol. 11, no. 110, pp. 81–90, 2024, doi:10.19101/IJATEE.2023.10102081. [Online]. Available: https://doi.org/10.19101/IJATEE.2023.10102081

[22] X. Qin, X. Liu, L. Xia, Q. Luo, and C. Zhang, "Multimodal ultrasound deep learning to detect fibrosis in early chronic kidney disease," *Renal Failure*, vol. 46, no. 2, pp. 1–10, 2024, doi:10.1080/0886022X.2024.2417740. [Online]. Available: https://doi.org/10.1080/0886022X.2024.2417740

[23] S. B S, "Kidney disease classification using deep learning," *International Journal of Scientific Research in Engineering and Management*, vol. 9, no. 5, pp. 1–6, 2025, doi:10.55041/ijsrem48296. [Online]. Available: https://doi.org/10.55041/ijsrem48296

[24] P. Liang, J. Yang, W. Wang, G. Yuan, M. Han,

Q. Zhang, and Z. Li, "Deep learning identifies intelligible predictors of poor prognosis in chronic kidney disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3500–3510, 2023, doi:10.1109/JBHI.2023.3266587. [Online]. Available: https://doi.org/10.1109/JBHI.2023.3266587

[25] Y. Tan, Y. Ma, S. Rao, and X. Sun, "Performance of deep learning for detection of chronic kidney disease from retinal fundus photographs: A systematic review and meta-analysis," *European Journal of Ophthalmology*, vol. 34, no. 2, pp. 512–520, 2024, doi:10.1177/11206721231199848. [Online]. Available: https://doi.org/10.1177/11206721231199848

[26] M. N. H. Chowdhury, M. B. I. Reaz, S. H. M. Ali, M. L. Crespo, S. Ahmad, G. M. Salim, F. Haque, L. G. G. Ordóñez, M. J. Islam, T. M. Mahdee, K. S. Zaman, M. S. K. Hemel, and M. A. S. Bhuiyan, "Deep learning for early detection of chronic kidney disease stages in diabetes patients: A tabnet approach," *Artificial Intelligence in Medicine*, vol. 166, p. Article 103153, 2025, doi:10.1016/j.artmed.2025.103153. [Online]. Available: https://doi.org/10.1016/j.artmed.2025.103153

[27] L. Holmstrom, M. Christensen, N. Yuan, J. W. Hughes, J. Theurer, M. Jujjavarapu, P. Fatehi, A. Kwan, R. K. Sandhu, J. Ebinger, S. Cheng, J. Zou, S. S. Chugh, and D. Ouyang, "Deep learning-based electrocardiographic screening for chronic kidney disease," *Communications Medicine*, vol. 3, no. 1, p. Article 78, 2023, doi:10.1038/s43856-023-00278-w. [Online]. Available: https://doi.org/10.1038/s43856-023-00278-w

[28] R. H. Aswathy, P. Suresh, M. Y. Sikkandar, S. Abdel-Khalek, H. Alhumyani, R. A. Saeed, and R. F. Mansour, "Optimized tuned deep learning model for chronic kidney disease classification," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2453–2468, 2022, doi:10.32604/cmc.2022.019790. [Online]. Available: https://doi.org/10.32604/cmc.2022.019790

[29] M. N. Hossain, E. Bhuiyan, M. B. A. Miah, T. A. Sifat, Z. Muhammad, and M. F. A. Masud, "Detection and classification of kidney disease from ct images: An automated deep learning approach," *Technologies*, vol. 13, no. 11, p. 508, 2025, doi:10.3390/technologies13110508. [Online]. Available: https://doi.org/10.3390/technologies13110508

[30] I. I. Ayogu, C. F. Daniel, B. A. Ayogu, J. N. Odii, C. L. Okpalla, and E. C. Nwokorie, "Investigation of ensembles of deep learning models for improved chronic kidney diseases detection in ct scan images," *Franklin Open*, vol. 11, p. Article 100298, 2025, doi:10.1016/j.fraope.2025.100298. [Online]. Available: https://doi.org/10.1016/j.fraope.2025.100298

[31] P. K. Rao, S. Chatterjee, K. Nagaraju, S. B. Khan, A. Almusharraf, and A. I. Alharbi, "Fusion of graph and tabular deep learning models for predicting chronic kidney disease," *Diagnostics*, vol. 13, no. 12, p. 1981, 2023, doi:10.3390/diagnostics13121981. [Online]. Available: https://doi.org/10.3390/diagnostics13121981

[32] A. S. Mahmoud, O. Lamouchi, and S. Belghith, "Advancements in machine learning and deep learning for early diagnosis of chronic kidney diseases: A comprehensive review," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 1–15, 2024, doi:10.58496/BJML/2024/015. [Online]. Available: https://doi.org/10.58496/BJML/2024/015

[33] A. Yumashev, P. Udayakumar, S. N. S. V. S. C. Ramesh, E. L. Lydia, and K. V. Kumar, "Role of rough neutrosophic attribute reduction with deep learning-based enhanced kidney disease diagnosis," *International Journal of Neutrosophic Science*, vol. 25, no. 1, 2025, doi:10.54216/IJNS.250126.

[34] K. Kumar, M. Pradeepa, M. Mahdal, S. Verma, M. V. RajaRao, and J. V. N. Ramesh, "A deep learning approach for kidney disease recognition and prediction through image processing," *Applied Sciences*, vol. 13, no. 6, p. 3621, 2023, doi:10.3390/app13063621. [Online]. Available: https://doi.org/10.3390/app13063621

[35] B. Dharaneesh, S. Sujan, and Hemalatha, "Deep learning architectures for kidney disease classification," *International Journal of Scientific Research in Engineering and Management*, vol. 8, no. 8, pp. 1–6, 2024, doi:10.55041/ijsrem37291. [Online]. Available: https://doi.org/10.55041/ijsrem37291

[36] D. Saif, A. M. Sarhan, and N. M. Elshennawy, "Deep-kidney: an effective deep learning framework for chronic kidney disease prediction," *Health Information Science and Systems*, vol. 12, no. 1, 2024, doi:10.1007/s13755-023-00261-8.

[37] M. Zhang, Z. Ye, E. Yuan, X. Lv, Y. Zhang, Y. Tan, C. Xia, J. Tang, J. Huang, and Z. Li, "Imaging-based deep learning in kidney diseases: Recent progress and future prospects," *Insights into Imaging*, vol. 15, no. 1, p. Article 36, 2024, doi:10.1186/s13244-024-01636-5. [Online]. Available: https://doi.org/10.1186/s13244-024-01636-5

[38] F. Mallamaci and G. Tripepi, "Risk factors of chronic kidney disease progression: Between old and new concepts," *Journal of Clinical Medicine*, vol. 13, no. 3, p. 678, 2024, doi:10.3390/jcm13030678. [Online]. Available: https://doi.org/10.3390/jcm13030678

[39] Ákos Géza Pethő, M. Tapolyai, Éva Csongrádi, and P. Orosz, "Management of chronic kidney disease: The current novel and forgotten therapies," *Journal of Clinical and Translational Endocrinology*, vol. 36, p. 100354, 2024, doi:10.1016/j.jcte.2024.100354.

[40] F. G. Scurt, M. J. Ganz, C. Herzog, K. Bose, P. R. Mertens, and C. Chatzikyrkou, "Association of metabolic syndrome and chronic kidney disease," *Obesity Reviews*, vol. 25, no. 1, p. N/A, 2024, doi:10.1111/obr.13649.

[41] K. Tsuji, N. Uchida, H. Nakanoh, K. Fukushima, S. Haraguchi, S. Kitamura, and J. Wada, "The gut–kidney axis in chronic kidney diseases," *Diagnostics*, vol. 15, no. 1, p. 21, 2025, doi:10.3390/diagnostics15010021.

[42] P. Tao, C. W. Chien, C. Liu, J. Zheng, D. Sun, J. Zeng, Q. Song, Y. Liu, T. H. Tung, and L. Kang, "Diabetes mellitus is a risk factor for incident chronic kidney disease: A nationwide cohort study," *Heliyon*, vol. 10, no. 7, 2024, doi:10.1016/j.heliyon.2024.e28780.

[43] N. Pradhan, J. Kerner, L. A. Campos, and M. Dobre, "Personalized nutrition in chronic kidney disease," *Biomedicines*, vol. 13, no. 3, p. 647, 2025, doi:10.3390/biomedicines13030647.

[44] P. I. Georgianos and R. Agarwal, "Hypertension in chronic kidney disease - treatment standard 2023,"

Nephrology Dialysis Transplantation, vol. 38, no. 12, p. N/A, 2023, doi:10.1093/ndt/gfad118.

[45] L. Cirillo, L. D. Chiara, S. Innocenti, C. Errichiello, P. Romagnani, and F. Becherucci, "Chronic kidney disease in children: an update," *Clinical Kidney Journal*, vol. 16, no. 10, p. N/A, 2023, doi:10.1093/ckj/sfad097.

[46] K. Badura, J. Janc, J. Wąsik, S. Gnitecki, S. Skwira, E. Młynarska, J. Rysz, and B. Franczyk, "Anemia of chronic kidney disease—a narrative review of its pathophysiology, diagnosis, and management," *Biomedicines*, vol. 12, no. 6, p. 1191, 2024, doi:10.3390/biomedicines12061191.

[47] S. Maringhini and C. Zoccali, "Chronic kidney disease progression—a challenge," *Biomedicines*, vol. 12, no. 10, p. 2203, 2024, doi:10.3390/biomedicines12102203.

[48] Y. Li and R. Padman, "Enhancing end-stage renal disease outcome prediction: a multisourced data-driven approach," *Journal of the American Medical Informatics Association*, vol. 33, no. 1, pp. 1–12, 2026, doi:10.1093/jamia/ocaf118.

[49] M. S. A. Huda, E. Kanon, M. S. K. Pappo, M. A. Ali, and N. Ahmed, "Nefroai: An explainable and real-time framework for predicting chronic kidney disease using diverse machine learning models and different feature selection techniques," *IEEE Access*, vol. 14, pp. 1–15, 2026, doi: 10.1109/ACCESS.2025.3649006.

[50] Y. Zhang, Y. Hou, T. Qiu, Y. Zhuang, K. Chen, W. Ling, Y. Luo, and J. Lin, "Deep learning in renal ultrasound: applications, challenges, and future outlook," *Frontiers in Oncology*, vol. 15, p. 1730628, 2026, doi:10.3389/fonc.2025.1730628.

[51] S. Sudharson and P. Kokil, "Automated diagnosis tool for clinical ultrasound image analysis in kidney disease identification," *Biomedical Signal Processing and Control*, vol. 111, p. 108345, 2026, doi:10.1016/j.bspc.2025.108345.

[52] V. P. Samuel and G. Annasamy, "Ultrasensitive detection of creatinine using deep learning-integrated graphene oxide gold nanocomposites sers sensor," *IEEE Sensors Letters*, vol. 10, no. 1, pp. 1–4, 2026, doi:10.1109/LSENS.2025.3642276.

[53] P. I. Nainggolan, F. Lubis, M. F. SyahPutra, T. G. M.R, A. Lumbangaol, and R. Sulaiman, "Advancing kidney disease diagnosis with yolov9 deep learning: A comprehensive classification approach," pp. 1–6, 2026, doi:10.1109/iccai65301.2025.11279427.

[54] S. S and V. A, "A hybrid deep learning approach based on optimized feature selection on environmental multi-disease predictive models," *International Journal of Cognitive Computing in Engineering*, vol. 7, pp. 100 150–100 165, 2026, doi:10.1016/j.ijcce.2025.11.006.

[55] S. Samenezhad and D. Rafighi, "The role of artificial intelligence in advancing urologic care: From diagnostics to therapeutics," *Surgery in Practice and Science*, vol. 24, p. 100322, 2026, doi:10.1016/j.sipas.2025.100322.

[56] W. Cheungpasitporn, A. Athavale, L. Ghazi, K. B. Kashani, T. Colicchio, J. L. Koyner, J. Chen, J. H. Ix, G. Nadkarni, and J. A. Neyra, "Transforming nephrology through artificial intelligence: A state-of-the-art roadmap for clinical integration," *Clinical Kidney Journal*, vol. 19, no. 1, pp. 1–15, 2026, doi:10.1093/ckj/sfag004.

[57] M. N. Hossain, E. Bhuiyan, M. B. A. Miah, S. Awang, M. S. Hossain, and T. A. Sifat, "Classification of kidney disease from ct images: A deep learning-based comparative analysis," pp. 1–6, 2026, doi:10.1109/icsecs65227.2025.11279151.

[58] R. O. Ogundokun, P. A. Owolawi, and E. A. van Wyk, "A comparative analysis of deep transfer learning and traditional handcrafted feature approaches for kidney disease classification using random forests," vol. 2667, pp. 205–218, 2026.

[59] S. M, T. M, E. R. G, and A. M, "A chronic kidney disease prediction system based on internet of things using walrus optimized deep learning technique," *Informatics for Health and Social Care*, pp. 1–15, 2026, doi:10.1080/17538157.2025.2610695.

[60] I. Sargunan, S. K. Krishnamoorthy, K. Govindharaju, and V. Cheemalapati, "Topk-kiu-net and squeeze residual fraction network for kidney tumor segmentation and detection," *Biomedical Signal Processing and Control*, vol. 113, p. 109116, 2026, doi:10.1016/j.bspc.2025.109116.

[61] M. Abniki, M. Amirdosara, and M. Zangi, "Ai-driven innovations in intensive care nephrology; bridging intensive care and kidney diseases," *Journal of Nephropharmacology*, vol. 15, no. 1, 2026, doi:10.34172/npj.2025.12820.

[62] Y. Chen, C. Chang, Y. Yang, X. Zhang, P. Chao, C. Chen, and C. Lu, "Rapid diagnosis of chronic kidney disease via raman spectroscopy combined with a lightweight mixture of experts method," *Journal of Raman Spectroscopy*, vol. 57, no. 1, 2026, doi:10.1002/jrs.70028.

[63] J. Eda, T. Mohaidat, S. I. Niloy, Z. Niu, M. R. K. Khan, and K. Khalil, "Artificial intelligence in healthcare: A comprehensive survey on ai-driven diagnosis and patient monitoring," *Applied Soft Computing*, vol. 190, p. 114429, 2026 doi:10.1016/j.asoc.2025.114429.

[64] S. C. Kunjumohamad and S. Govindaraj, "Nlp-driven federated learning framework for multi-organ rare disease prediction," *Iran Journal of Computer Science*, vol. 9, no. 1, 2026, doi:10.1007/s42044-025-00375-0.

[65] W. Dai, R. Liu, Z. Wu, T. Wu, M. Wang, J. Zhou, Y. Yuan, and J. Liu, "Exploiting scale-variant attention for segmenting small medical objects," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2026, doi:10.1109/TNNLS.2025.3645355.

[66] Z. A. Khan, M. Waqar, H. U. Khan, N. I. Chaudhary, A. T. Khan, I. Ishtiaq, F. A. Khan, and M. A. Z. Raja, "Fine-tuned deep transfer learning: An effective strategy for the accurate chronic kidney disease classification," *PeerJ Computer Science*, vol. 11, p. e2800, 2025, doi:10.7717/peerj-cs.2800. [Online]. Available: https://doi.org/10.7717/peerj-cs.2800

Case-Based Reasoning Model for Predicting the Malaria Cases

Konan N'gatta Aimé Kouassi[1], Koffi Kouakou Ive Arsene[2], Gooré Bi Tra[3]
Dept. Data Science and Artificial Intelligence Laboratory (LASDIA),
Institut National Polytechnique Felix, HOUPHOUET Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire[1,3]
Dept. Computer Science Research and Telecommunication (LARIT),
Institut National Polytechnique Felix, HOUPHOUET Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire[2]

***Abstract*—Malaria remains a major public health issue in Ivory Coast, where the need for accurate and interpretable predictive models is critical for effective disease control. While most existing approaches prioritize predictive accuracy over interpretability, this study addresses the need for explainable models suitable for deployment in resource-limited public health settings. The study used a dataset covering multiple regions of Côte d'Ivoire over the period 2019–2023 and including climatic variables such as temperature, humidity, and precipitation. Seven conventional machine learning models (CatBoost, XGBR, RFR, DTR, SVM, KNN, and Linear Regression) were compared with three proposed Case-Based Reasoning variants (CBR_0, CBR_1, and CBR_2), which differ in their similarity-weighting strategies and correction constant. The results show that CBR_2 achieved the best predictive performance, with RMSE = 0.081, MAE = 0.057, and $R^2 = 72.40\%$, followed by the Random Forest Regressor. A Wilcoxon signed-rank test confirmed a statistical significant difference of this permofance ($W = 18029$, $p = 1.02 \times 10^{-20}$). Beyond predictive accuracy, qualitative criteria including explainability, transparency, and adaptability were evaluated, further highlighting the superiority of CBR_2 over conventional black-box models. These findings highlight the potential of Case-Based Reasoning for epidemiological forecasting and decision support in malaria control.**

Keywords*—Malaria prediction; case-based reasoning; machine learning; epidemiological forecasting; explainable artificial intelligence

I. Introduction

Malaria remains a major public health problem, particularly in sub-Saharan Africa, where it continues to cause thousands of deaths each year despite considerable efforts in prevention and treatment [1]. According to the official WHO report dated December 4, 2025, 282 million cases of malaria were reported in 80 countries worldwide, resulting in 610,000 deaths. 95% of the cases and 95% of the deaths were recorded in the WHO African Region. This disease has disastrous consequences in Ivory Coast, where it is the leading cause of death among children under 5 and the main reason for medical consultations at health facilities across the country.

According to Dr. Tedros Adhanom Ghebreyesus, Executive Director of the WHO since 2017: "Malaria elimination is achievable in many countries, provided that sustained investments are made and efforts are stepped up to reach those most at risk". Modern tools, such as artificial intelligence models, are therefore essential to ensure better monitoring of the disease and to support effective malaria control and elimination efforts.

A major challenge is accurately predicting malaria cases to optimize the allocation of often limited medical resources and identify high-risk areas [2]–[6] Such predictions also support the development of early warning systems for timely intervention and improved disease control [7], [8]

Although several machine learning (ML) models have been proposed to predict malaria incidence [2]–[4], these approaches share common limitations. First, most of them require large volumes of training data, which are rarely available in resource-limited settings [9]. Second, and more critically, they predominantly rely on black-box architectures whose internal decision logic is opaque, making it difficult for health professionals and policy-makers to trust and act upon their outputs [10], [11]. Despite significant advances in Explainable Artificial Intelligence (XAI), a significant gap remains between the development of explainable methods and their actual applicability in operational contexts [12], most current XAI approaches rely on post-hoc techniques (e.g., SHAP) that explain predictions after the fact without guaranteeing that these explanations accurately reflect the model's internal workings [10]. Crucially, no existing study in the literature has incorporated an inherently explainable model such as Case-Based Reasoning (CBR) to predict the number of malaria cases in a given locality.

Furthermore, according to Caro-Martinez et al. [12], in healthcare it is crucial to understand how an artificial intelligence model generates its predictions, as this is essential for ensuring its reliable and trustworthy use. It not only helps build user confidence but also facilitates scientific validation, error detection, and continuous improvement of the model. In addition, Pradeep et al. [10] show that in many sensitive fields such as healthcare, finance, and the justice system, automated decisions must be transparently justified in order to meet ethical and regulatory requirements and facilitate their integration into decision-support systems. Thus, there is a lack of inherently explainable models capable of balancing predictive performance with transparency [10].

With this in mind, this study focuses on Case-Based Reasoning (CBR) as a natively explainable paradigm for malaria case prediction, examining both its theoretical foundations and its practical applicability in a real-world epidemiological context.

The CBR approach appears to be a particularly relevant alternative for addressing the limitations of current XAI approaches [11]. Unlike post-hoc methods, CBR is inherently

explainable and transparent, as it relies on a reasoning mechanism based on similar past cases [11]. Each prediction is thus supported by references to situations that have already been observed, providing a natural way to answer the questions "why" and "how" a decision was made. This form of explanation, based on concrete examples, is more intuitive and accessible to non-expert users, particularly business decision-makers [11]. In short, by combining traceability, transparency, and ease of explanation, CBR remains a promising approach for designing artificial intelligence systems that are understandable, high-performing, and operational in critical fields such as healthcare [10], [13].

In practical applications, CBR offers several advantages, notably its inherent interpretability: each prediction is justified by reference to previous cases, making the model intuitive for non-technical users [13]. Furthermore, thanks to its adaptability, the system can be expanded by adding new cases to the knowledge base without requiring complete retraining [9]. Finally, CBR performs well with small datasets, unlike ML models that require large amounts of data [9]. Despite these advantages, no existing study in the literature incorporates a CBR approach to predict the number of malaria cases in a given locality.

However, several challenges and limitations associated with the CBR approach must be acknowledged. In particular, prediction accuracy depends heavily on how the case base is organized [11], [14]. Furthermore, determining an appropriate similarity metric for heterogeneous data (numeric, categorical, and temporal) remains a challenge [11], [15]. Additionally, traditional CBR does not always account for how case characteristics evolve over time [16], [17].

The objective of this study is therefore to propose an enhanced CBR-based approach that is explainable, transparent, and effective to improve the prediction of malaria case counts at the locality level in Côte d'Ivoire, thereby addressing both the predictive limitations of classical CBR and the opacity of existing ML-based forecasting models.

In the remainder of this study, Section II presents a summary of previous work on predicting the number of malaria cases. Section III outlines the classical CBR approach. Section IV describe the proposed CBR approach. Section V presents the experimental comparison with classical CBR and ML models using data from Côte d'Ivoire. The results are discussed in Section VI. Section VII concludes the study with a summary and outlook for future work.

II. Related Work

This section provides background on malaria case prediction using data analysis and machine learning approaches, and introduces the classical CBR approach.

A. Statistical Models

Classical statistical approaches have been widely applied to malaria time series forecasting. Linear regression and ARIMA models have been used to model temporal trends in case counts [18]–[20], while their seasonal extensions (SARIMA) have been adopted to capture recurrent epidemic cycles [21]. These models have demonstrated their ability to reflect seasonal patterns and weather-driven fluctuations in malaria incidence [3], [22].

However, statistical models suffer from several structural limitations. First, they rely on restrictive assumptions linearity, stationarity, and variable independence that are rarely met in real epidemiological data [23]. Second, and more critically, they are inherently unable to capture the nonlinear relationships and complex interactions among the climatic, environmental, and socioeconomic determinants of malaria transmission [22], [24], [25]. These shortcomings motivated the shift toward machine learning approaches, discussed in the following section.

B. Machine Learning Approaches

Several ML models have been developed to predict malaria case counts, with the aim of anticipating disease progression and facilitating more efficient allocation of medical resources [2]. The studies reviewed in this section are organized around three main axes: 1) the types of ML models employed, 2) the input features and data sources used, and 3) the limitations identified.

1) Machine learning models for malaria prediction: A wide range of ML algorithms has been applied to malaria case prediction. Ensemble methods have received considerable attention: Random Forest models have been used by [18] and [26] to capture non-linear relationships between environmental variables and malaria incidence, achieving competitive accuracy across different geographic settings. Gradient Boosting methods, including XGBoost and LightGBM, have been explored by [27], who demonstrated their superiority over single-tree models in terms of predictive performance on imbalanced epidemiological datasets. Kernel-based methods such as Support Vector Machines (SVM) have also been applied [28], particularly in contexts where the training dataset is limited in size. More recently, deep learning architectures have emerged in related infectious disease prediction tasks: for instance, [29] proposed CONV-FAN-POX, a convolutional deep learning framework for monkeypox detection, illustrating the growing trend toward deep feature extraction in epidemiological surveillance.

Beyond individual models, ML approaches have been shown to serve broader public health objectives. In [5], [6], and [30], ML models have helped public health authorities develop dynamic budgets, identify high-risk areas, and implement early warning systems.

2) Limitations of existing studies: Despite these advances, the existing literature reveals several recurring and significant limitations.

a) High data requirements: Most ML models, particularly deep learning architectures and ensemble methods, require large volumes of labeled training data [1], [9]. This constitutes a critical bottleneck in sub-Saharan Africa, where epidemiological surveillance systems are often incomplete or insufficiently digitized [4].

b) Lack of explainability and transparency: The models used in the reviewed studies are predominantly *black-box* models that do not provide intelligible justifications for their predictions [11], [28]. This opacity is particularly problematic

in healthcare contexts, where clinicians and policy-makers need to understand and trust the reasoning behind a prediction [12], [31]. Although some studies apply post-hoc explanation techniques such as SHAP or LIME, these methods do not guarantee that the explanation faithfully reflects the model's internal logic [10] and are often difficult to interpret for non-expert users [11].

c) Poor adaptability to temporal dynamics: Most ML models reviewed are trained statically and do not incorporate mechanisms for continuous updating as new epidemiological data become available [16], limiting their reliability for operational forecasting.

d) Limited geographic transferability: Many models were developed and validated in specific geographic contexts, and their performance in different settings such as Côte d'Ivoire has rarely been assessed [2], [3].

e) Absence of inherently explainable approaches: Crucially, no study in the reviewed literature has proposed an inherently explainable model for malaria case prediction. The CBR paradigm, which is natively interpretable and well suited to small datasets, has not been explored in this context, a gap that the present study aims to fill.

C. Introduction to Case-Based Reasoning

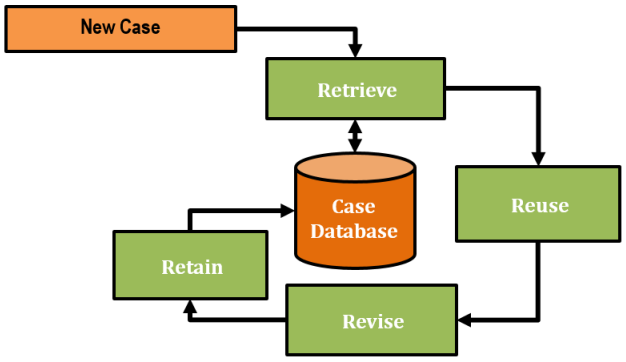


Fig. 1. The CBR life cycle.

Case-Based Reasoning (CBR) is an artificial intelligence method that solves new problems by drawing on concrete past experiences, known as cases, which contain a description of the problem, the solution applied, and the results obtained. The most similar cases are retrieved from a case base and their solutions are adapted to address the new problem. This enables intuitive and easily explainable reasoning, analogous to the way humans solve problems by drawing on past experiences. As shown in Fig. 1, the CBR life cycle comprises five sequential tasks [32]:

- Define: a new problem is described in order to compare it to past problems stored in the case base.
- Retrieve: using a similarity measure, the most similar problems in the case base are searched and selected.
- Reuse: knowledge related to the selected cases is reused to propose an initial solution.
- Revise: if the initial solution is not suitable, it is revised to make it more appropriate.
- Retain: the new problem and its associated solution are stored in the case base.

Major challenges associated with this approach include structuring the case database, selecting appropriate similarity metrics, and accounting for changes in the target variable over time (as discussed in Section I). To address these shortcomings, an improved CBR approach for predicting the number of malaria cases is proposed in the following section.

III. Proposed CBR Approach

This section presents the enhanced Case-Based Reasoning (CBR) approach proposed for predicting the number of malaria cases. The classical CBR cycle Retrieve, Reuse, Revise, Retain is extended through two key contributions: 1) a regional clustering mechanism that restricts similarity search to epidemiologically homogeneous groups, and 2) an error correction constant that compensates for temporal drift in case characteristics. The rationale for each design choice is provided alongside its mathematical formulation. Fig. 2 provides an overview of the proposed architecture.

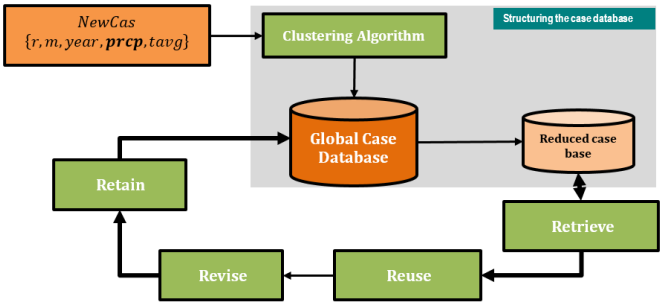


Fig. 2. Illustration of the proposed CBR approach.

A. Notation and Preliminary Definitions

To ensure clarity throughout this section, Table I summarizes the main symbols used in the proposed model.

TABLE I. Summary of Mathematical Notation

Symbol	Definition
Ω	Global case base (all historical cases)
ϕ	Reduced case base (regional cluster)
Cas_t	Target case (new case to predict)
Cas_p^i	i-th previous case in ϕ
att_i^t, att_i^p	Value of attribute i for target and previous case
$sim_i(\cdot,\cdot)$	Local similarity for attribute i
$Sim(\cdot,\cdot)$	Global (weighted) similarity
w_i^*	Normalized weight of attribute i
W_j	Prediction weight of the j-th neighbor
m	Number of nearest neighbors retrieved
k	Total number of attributes
n	Number of instances in Ω
cst	Temporal correction constant
$Pred(\text{Cas}_t)$	Predicted number of malaria cases

B. Case Structuring

In a CBR system, a *case* encapsulates a past experience comprising three elements: a problem description, the solution

applied, and the outcome obtained [32]. Accurate and consistent case representation is a prerequisite for reliable similarity computation and solution adaptation. Three properties are particularly critical in this context:

- Retrieval relevance and solution adaptation: Homogeneous case representation improves similarity computation and enables the retrieval of the most contextually relevant historical cases [9].
- Explainability: A clear and structured case representation makes the system's reasoning transparent and its predictions interpretable by non-expert users [9].
- Integration of domain knowledge: A well-defined structure allows epidemiological knowledge (e.g., seasonality, regional characteristics) to be embedded directly into the case representation [11].

Each past and new case is represented as a feature vector containing six attributes selected on the basis of their established influence on malaria transmission [5]:

$$\text{Cas} = \{r,\ m,\ year,\ prcp,\ tavg,\ hmdt\} \tag{1}$$

where, r denotes the geographic region, m the month, $year$ the year, $prcp$ the precipitation, $tavg$ the average temperature, and $hmdt$ the humidity. The target variable to be predicted is the number of malaria cases associated with each case. The choice of these features reflects the climatic and spatio-temporal determinants of malaria incidence: temperature and precipitation directly influence mosquito breeding cycles, humidity affects vector survival, while region and month capture spatial heterogeneity and seasonal periodicity respectively.

C. Case Base Structuring via Regional Clustering

In a naive CBR system, similarity is computed against the entire case base Ω, which introduces two problems: 1) irrelevant cases from geographically distant regions pollute the similarity ranking, and 2) local epidemiological characteristics specific to each region such as hydrography, vegetation density, and population structure are not accounted for. To address these issues, Ω is partitioned into regional clusters $\phi_1, \phi_2, \ldots, \phi_k$ using the attribute r (region) as the grouping criterion. During retrieval, only the cluster ϕ_i corresponding to the target case's region is searched, thereby reducing the search space and improving the contextual relevance of retrieved cases.

Formally, the clustering operation partitions Ω as follows:

$$\Omega = \bigcup_{i=1}^{k} \phi_i, \quad \phi_i \cap \phi_j = \emptyset \ \forall i \neq j \tag{2}$$

where, each cluster ϕ_i groups all cases belonging to region r_i. Algorithm 1 formalizes this partitioning procedure, and Fig. 3 illustrates the resulting case base structure.

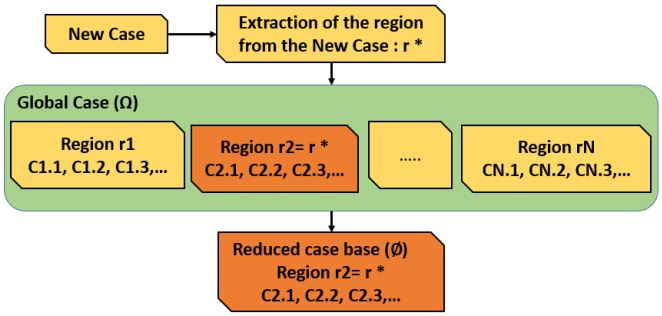


Fig. 3. Process of case base reduction via regional clustering.

Algorithm 1 Case Base Partitioning by Region

Require: $\Omega = \{(X_i, Y_i)\}_{i=1}^{N}$: Global case base
Require: $R = \{r_1, r_2, \ldots, r_p\}$: Set of regions in Ω
Ensure: Regional clusters $\phi_1, \phi_2, \ldots, \phi_k$
  Initialize an empty set of clusters $\phi \leftarrow \emptyset$
  for each region $r_i \in R$ **do**
    if cluster ϕ_k associated with r_i exists **then**
      Add case to ϕ_k
    else
      Create new cluster $\phi_k \leftarrow \{r_i\}$
    end if
  end for
  return $\phi_1, \phi_2, \ldots, \phi_k$

D. Similarity Computation

Similarity computation is the core mechanism of CBR retrieval. Given a target case Cas_t and a previous case $\text{Cas}_p^i \in \phi$, the overall similarity $Sim(\text{Cas}_t, \text{Cas}_p^i)$ is obtained through a two-step process: 1) computation of a *local similarity* for each attribute, and 2) aggregation into a *global similarity* via a weighted sum. The local similarity metric is chosen according to the attribute type categorical, numerical, or cyclic as detailed below.

1) Local similarity metrics:

a) Categorical attribute Region (r): Since region is a nominal attribute with no inherent ordering, a binary metric is used [33]. Two cases are fully similar if they belong to the same region, and fully dissimilar otherwise.

$$sim(r_t, r_p^i) = \begin{cases} 1 & \text{if } r_t = r_p^i \\ 0 & \text{if } r_t \neq r_p^i \end{cases} \tag{3}$$

b) Numerical attribute year ($year$): A normalized numerical distance is used to assign higher similarity to temporally proximate cases, reflecting the assumption that recent epidemiological patterns are more relevant than distant ones [34]:

$$sim(y_t, y_p^i) = 1 - \frac{|y_t - y_p^i|}{\max_{p \in \phi} |y_t - y_p^i|} \tag{4}$$

c) Cyclic attribute month (m): Month is a periodic variable: the distance between December ($m = 12$) and January ($m = 1$) should be 1, not 11. A circular distance is therefore adopted to account for the periodic nature of the annual cycle:

$$sim(m_t, m_p^i) = 1 - \frac{\min(|m_t - m_p^i|,\ 12 - |m_t - m_p^i|)}{6} \quad (5)$$

The denominator 6 corresponds to the maximum possible circular distance between two months, ensuring the similarity is bounded in $[0, 1]$.

d) Numerical attributes temperature ($tavg$), Precipitation ($prcp$), Humidity ($hmdt$): These three continuous attributes are normalized to $[0, 1]$ via Min-Max scaling prior to similarity computation. The normalized absolute difference directly yields a bounded similarity in $[0, 1]$ [33]:

$$sim(tavg_t, tavg_p^i) = 1 - |tavg_t - tavg_p^i| \quad (6)$$
$$sim(prcp_t, prcp_p^i) = 1 - |prcp_t - prcp_p^i| \quad (7)$$
$$sim(hmdt_t, hmdt_p^i) = 1 - |hmdt_t - hmdt_p^i| \quad (8)$$

The same functional form is applied to these three attributes because they share the same scale after normalization, and smaller absolute differences reflect greater contextual similarity for malaria transmission conditions.

2) Global similarity: The global similarity between the target case Cas_t and a previous case Cas_p is defined as the weighted sum of all local similarities:

$$Sim(\text{Cas}_t, \text{Cas}_p) = \sum_{i=1}^{k} w_i^* \times sim_i(att_i^t, att_i^p) \quad (9)$$

In the absence of domain-specific evidence prioritizing one attribute over another, equal weights are assigned to all k attributes to avoid the introduction of subjective bias:

$$w_i = 1, \quad \forall i \in \{1, 2, \ldots, k\} \quad (10)$$

To ensure that the global similarity remains in $[0, 1]$, the weights are normalized:

$$w_i^* = \frac{w_i}{\sum_{j=1}^{k} w_j} = \frac{1}{k}, \quad \forall i \in \{1, 2, \ldots, k\} \quad (11)$$

It follows that $\sum_{i=1}^{k} w_i^* = 1$, which guarantees $Sim(\text{Cas}_t, \text{Cas}_p) \in [0, 1]$ given that each $sim_i \in [0, 1]$.

E. Prediction of Malaria Cases

Once the m most similar cases $\{\text{Cas}_j^p\}_{j=1}^m$ are retrieved from ϕ, the predicted number of malaria cases is estimated as a similarity-weighted average of the target values of the retrieved neighbors:

$$Pred(\text{Cas}_t) = \sum_{j=1}^{m} W_j \times Pred(\text{Cas}_j^p) \quad (12)$$

where, the prediction weight W_j of each neighbor is proportional to its global similarity to the target case:

$$W_j = \frac{Sim(\text{Cas}_t, \text{Cas}_j^p)}{\sum_{j=1}^{m} Sim(\text{Cas}_t, \text{Cas}_j^p)} \quad (13)$$

This weighted averaging scheme assigns greater influence to the most similar historical cases, while ensuring that $\sum_{j=1}^{m} W_j = 1$, so the prediction remains within the range of observed values.

F. Temporal Correction Constant

A fundamental limitation of classical CBR is the implicit stationarity assumption: the relationship between features and the target variable is assumed to remain constant over time. In the malaria context, this assumption is violated due to evolving climatic conditions, changes in healthcare interventions, and long-term trends in case counts. To compensate for this temporal drift, a correction constant cst is introduced, estimated as the mean absolute year-on-year variation in malaria case counts across all instances in Ω:

$$cst = \frac{1}{n} \sum_{i=1}^{n} \left| \text{NbreCas}_i^l - \text{NbreCas}_i^{l+1} \right| \quad (14)$$

where, NbreCas_i^l and NbreCas_i^{l+1} denote the number of malaria cases for instance i in years l and $l + 1$ respectively. The corrected prediction is then:

$$Pred(\text{Cas}_t) = \sum_{j=1}^{m} W_j \times Pred(\text{Cas}_j^p) + cst \quad (15)$$

The constant cst captures the average magnitude of inter-annual fluctuations, providing a simple, data-driven estimate of temporal drift that requires no additional parameters. However, since absolute differences are used, the direction of the trend is not encoded: the same correction is applied regardless of whether cases are increasing or decreasing. This limitation which may cause overestimation during sustained decline or underestimation during sustained growth is acknowledged and identified as a direction for future work (see Section VI).

G. Complete Model Formulation

Substituting Eq. (9), (13), and (14) into Eq. (15), the complete prediction equation of the proposed CBR model is:

$$Pred(\text{Cas}_t) = \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{k} w_i^* \cdot sim_i(att_i^t, att_i^p)}{\sum_{j=1}^{m} Sim(\text{Cas}_t, \text{Cas}_j^p)} \right) \times Pred(\text{Cas}_j^p) + \frac{1}{n} \sum_{i=1}^{n} \left| \text{NbreCas}_i^l - \text{NbreCas}_i^{l+1} \right| \tag{16}$$

This formulation integrates three components: 1) a *similarity-weighted retrieval* from a regionally clustered case base, 2) a *cyclic similarity metric* suited to the seasonal nature of malaria data, and 3) a *temporal correction* that reduces the bias introduced by epidemiological drift over time.

H. Proposed Model Algorithm

Algorithm 2 Enhanced CBR Model for Malaria Case Prediction

Require: $\Omega = \{(X_i, Y_i)\}_{i=1}^{N}$: Global case base
Require: $Train \subset \Omega$: Training set (X_i, Y_i)
Require: $Test \subset \Omega$: Test set (X_i)
Require: $\text{Cas}_t \in Test$: Target case $\{r, m, year, prcp, tavg, hmdt\}$
Require: $w_1^*, \ldots, w_k^*$: Normalized attribute weights
Require: $\{\phi_i\}_{i=1}^{p}$: Regional clusters
Require: m: Number of nearest neighbors
Step 1: Partition Ω into regional clusters $\phi_1, \ldots, \phi_k$ (Algorithm 1)
Step 2: Identify ϕ_i such that $\text{Cas}_t.r \in \phi_i$
Step 3: Apply Min-Max scaling to all attributes in ϕ_i
Step 4: Compute $sim_i(att_i^t, att_i^p)$ for each attribute and each $\text{Cas}_p \in \phi_i$
Step 5: Compute $Sim(\text{Cas}_t, \text{Cas}_p)$ via Eq. (9)
Step 6: Select the m most similar cases $\{\text{Cas}_j^p\}_{j=1}^{m}$
Step 7: Compute $Pred(\text{Cas}_t)$ via Eq. (16)
return $Pred(\text{Cas}_t)$

The proposed algorithm (Algorithm 2) consists of seven steps : regional partitioning of the case base, selection of the relevant cluster, attribute normalization, local and global similarity computation, retrieval of the m nearest neighbors, and final prediction incorporating the temporal correction constant.

IV. Experimentation

This section describes the dataset used to validate the proposed approach, the preprocessing steps applied, and the models and evaluation metrics employed.

A. Description of the Dataset

The data were collected from two institutions. Epidemiological data were obtained from the Ministry of Health of Côte d'Ivoire, which monitors malaria cases across all national health facilities. Monthly time series were constructed for each of the 33 health regions, reporting the number of severe malaria cases recorded between 2019 and 2023. Climate variables, including temperature and precipitation, were provided by the Airport, Aviation, and Meteorological Operations and Development Company (SODEXAM).

B. Preprocessing

Categorical variables (months and regions) were encoded using the `OneHotEncoder()` function from the `scikit-learn` library. The `MinMaxScaler()` method was applied to normalize the temperature variable, and the `RobustScaler()` function was used for the precipitation variable. These normalizations and encodings are necessary to provide machine learning models with a clean dataset that promotes effective learning.

- OneHotEncoder(): This function transforms categorical variables into binary numerical variables. Each category is converted into a separate column containing only the values 0 or 1.

$$x_i = \begin{cases} 1 & \text{if the category is present} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

- MinMaxScaler(): This method normalizes data by scaling all values into the range $[0, 1]$, preventing attributes with large numerical values from excessively influencing the model [1].

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{18}$$

where, x is the original value, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values respectively, and x' is the normalized value. This normalization is particularly suited to distance- and similarity-based methods.

C. Experimental Setup

The dataset was partitioned using a single temporal split, with 80% of the data (2019–2022) used for training and the remaining 20% (2023) reserved for testing. This chronological split was deliberately chosen over k-fold cross-validation for two reasons. First, malaria incidence data exhibits strong temporal dependencies and seasonal autocorrelation; a randomized fold assignment would mix past and future observations, introducing data leakage and producing overly optimistic performance estimates. Second, the evaluation is designed to simulate a realistic forecasting scenario in which a model trained on historical epidemiological data is used to predict future case counts for an unseen period. In this setting, preserving the temporal order of observations is methodologically essential. This approach is consistent with established practice in time-series forecasting evaluation. The features used were region, year, month, temperature, and precipitation. The objective was to predict the number of malaria cases.

Several supervised learning models were selected for comparison due to their widespread use in the literature and their recognized performance in regression tasks: CatBoost, XGBoost Regressor (XGBR), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and linear regression, as well as the three proposed CBR variants CBR_0, CBR_1, and CBR_2.

- CBR_2: This approach corresponds to the full proposed CBR method (see Section III). It incorporates the regional clustering algorithm (Algorithm 1), cyclic distance-based similarity, and the correction constant.
- CBR_1: This variant applies the same principles as CBR_2, using cyclic distance-based similarity, but without the clustering algorithm or the correction constant.
- CBR_0: This minimalist variant applies neither the clustering algorithm, the correction constant, nor the cyclic distance; numerical distance is used for the month attribute instead.

D. Evaluation Metrics

Two types of criteria were used to evaluate the models.

The quantitative criteria are:

- RMSE: the square root of the mean squared error; it heavily penalizes large errors and is sensitive to outliers [35].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

- MAE: the average of the absolute errors between actual and predicted values; it is easy to understand and penalizes all errors equally [36].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{20}$$

- $\mathbf{R^2}$: measures the proportion of variance explained by the model [36].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

The qualitative criteria are:

- Explainability (Expl): A model's ability to provide understandable justifications for its predictions [31].
- Transparency (Transp): A model's ability to make its internal workings accessible and interpretable [31].
- Adaptability (Adapt): A model's ability to incorporate new data without requiring a full retraining process [9].

V. Results

A. Training Set Performance

Table II reports the performance of ML models on the training set. KNN achieves perfect scores (RMSE = MAE = 0.00, $R^2 = 100\%$), followed by RFR (RMSE = 0.023, MAE = 0.016, $R^2 = 96.54\%$). These results suggest overfitting, particularly for KNN, as such scores are rarely reproducible on unseen data.

B. Test Set Performance

Table III presents the comparative evaluation on the test set, combining quantitative metrics (RMSE, MAE, R^2) and qualitative criteria (Explainability, Transparency, Adaptability).

CBR-based models achieve the best generalization performance. CBR_2 records the lowest errors (RMSE = 0.081, MAE = 0.057) and the highest coefficient of determination ($R^2 = 72.40\%$), followed by CBR_1 ($R^2 = 60.04\%$) and CBR_0 ($R^2 = 57.70\%$). This progressive improvement reflects the cumulative effect of the enhancements introduced at each stage. Among ML models, RFR and KNN show acceptable performance ($R^2 > 60\%$), while CatBoost, XGBR, DTR, SVM, and linear regression perform poorly ($R^2 < 40\%$), with linear regression recording the worst results (RMSE = 0.154, $R^2 = 2.18\%$).

Only CBR models (CBR_0, CBR_1, CBR_2) satisfy all three qualitative criteria ("Yes"). All ML models are classified "No", as their use requires specialized machine learning expertise, making them unsuitable for direct deployment by health personnel in Côte d'Ivoire.

C. Overall Ranking

Table IV provides a combined ranking based on both quantitative and qualitative scores. CBR_2 achieves the lowest total score (6), ranking first overall. RFR ranks second (score = 12), despite failing all qualitative criteria. CBR_1 and KNN share third place (score = 15), while linear regression ranks last (score = 34). These results confirm that CBR_2 achieves the best balance between predictive accuracy and operational transparency. Fig. 4 - 7 illustrate these results graphically.

D. Statistical Validation

A Wilcoxon signed-rank test was conducted to assess whether the superiority of CBR_2 over RFR is statistically significant. The test yields $W = 18{,}029$ and $p = 1.02 \times 10^{-20}$, well below the 0.05 threshold, confirming that the performance advantage of CBR_2 reflects a genuine difference in generalization capacity and is not an artefact of the temporal split.

TABLE II. Performance of ML Models on the Training Set

Model	RMSE	MAE	R^2 (%)
CatBoost	0.079	0.058	59.38
XGBR	0.098	0.073	37.37
RFR	0.023	0.016	96.54
DTR	0.099	0.075	36.16
SVM	0.102	0.061	32.35
KNN	0.000	0.000	100
Reg Lin	0.118	0.084	9.83

TABLE III. PERFORMANCE OF THE VARIOUS MODELS ON THE TEST SET

Quantitative Metrics				Qualitative Criteria		
Model	RMSE	MAE	R^2 (%)	Expl	Transp	Adapt
CBR_0	0.100	0.072	57.70	Yes	Yes	Yes
CBR_1	0.097	0.069	60.04	Yes	Yes	Yes
CBR_2	0.081	0.057	72.40	Yes	Yes	Yes
CatBoost	0.119	0.084	38.42	No	No	No
XGBR	0.140	0.099	15.84	No	No	No
RFR	0.094	0.067	62.22	No	No	No
DTR	0.138	0.099	18.03	No	No	No
SVM	0.140	0.092	15.45	No	No	No
KNN	0.096	0.068	60.47	No	No	No
Reg Lin	0.154	0.108	2.18	No	No	No

TABLE IV. MODEL RANKING ON THE TEST SET

Quantitative Ranking				Qualitative Ranking				
Model	RMSE	MAE	R^2	Expl	Transp	Adapt	Total	Rank
CBR_0	5	5	5	1	1	1	18	4
CBR_1	4	4	4	1	1	1	15	3
CBR_2	1	1	1	1	1	1	6	1
CatBoost	6	6	6	2	2	2	24	5
XGBR	8	8	8	2	2	2	30	7
RFR	2	2	2	2	2	2	12	2
DTR	7	8	7	2	2	2	28	6
SVM	8	7	9	2	2	2	30	7
KNN	3	3	3	2	2	2	15	3
Reg Lin	9	9	10	2	2	2	34	8

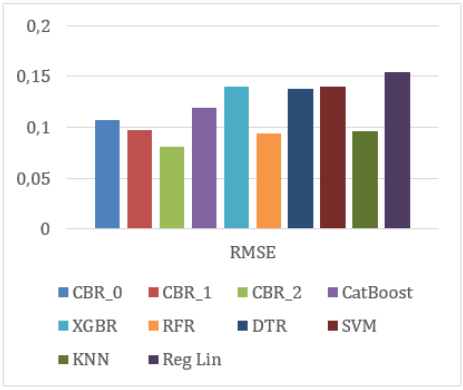


Fig. 4. Comparison of model RMSE on the test set.

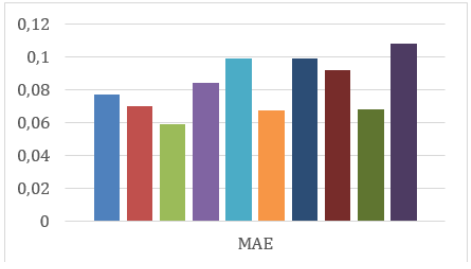


Fig. 5. Comparison of model MAE on the test set.

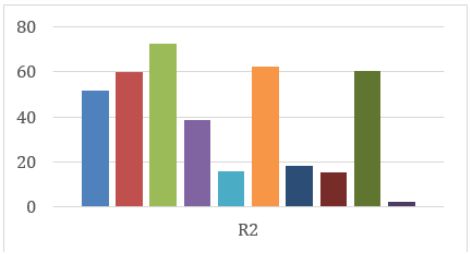


Fig. 6. Comparison of model R^2 on the test set.

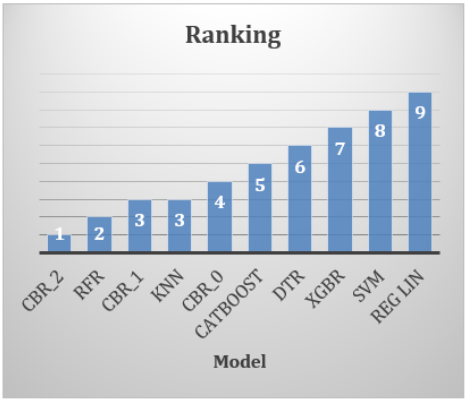


Fig. 7. Overall model ranking based on quantitative and qualitative criteria.

VI. DISCUSSION

A. Interpretation of CBR Performance

The superiority of CBR_2 over all other models results directly from three cumulative design improvements: the adoption of a cyclic similarity measure suited to seasonal epidemiological patterns, the optimization of the case base via clustering, and the introduction of a temporal correction constant. Each transition from CBR_0 to CBR_1, then to CBR_2 confirms that these choices jointly address the core challenge of malaria forecasting: capturing seasonality while maintaining generalization ability.

The significant performance drop observed for ML models between training and test sets reflects a classic overfitting phenomenon, particularly pronounced for KNN and RFR. This suggests that these models memorize training patterns rather than learn transferable relationships between climatic variables and malaria incidence, a limitation that CBR's instance-based reasoning naturally avoids.

B. The Added Value of Qualitative Criteria

Quantitative metrics alone are insufficient for evaluating models intended for operational deployment in public health. The qualitative evaluation reveals a structural divide: CBR models are natively explainable, transparent, and adaptable, whereas all ML models require specialized knowledge to interpret and maintain. In a context such as Côte d'Ivoire, where health personnel are not ML experts, this distinction is operationally decisive. A model that health workers can understand, question, and trust is more likely to influence decisions than a high-performing black box [12], [31].

C. Consistency with the Literature

The competitive performance of RFR ($R^2 = 62.22\%$) aligns with findings from Jinad et al., Lucas et al., and Monteiro et al., who identify Random Forest as one of the most effective ML models for malaria prediction [18], [26], [37]. However, the CBR_2 model proposed in this study surpasses RFR both quantitatively and qualitatively. The Wilcoxon signed-rank test ($W = 18{,}029$ and $p = 1.02 \times 10^{-20}$) confirms that this advantage is statistically significant, reflecting a genuine difference in generalization capacity.

D. Practical Implications

The CBR_2 model offers concrete value for public health decision-making. Its case-based reasoning mechanism enables health authorities to trace the historical cases that drove each prediction, facilitating early detection of high-risk periods and supporting timely preventive interventions. Its adaptability the ability to incorporate new cases without full retraining makes it particularly suited to operational surveillance settings where data are updated continuously and resources are limited.

E. Limitations of the Study

Despite these results, several limitations must be acknowledged.

1) Correction constant design: The error correction constant quantifies historical fluctuations independently of trend direction. It may therefore lead to overestimation during sustained decline periods or underestimation during sustained growth phases.

2) Temporal validation: The evaluation relies on a single 80/20 temporal split. Although the Wilcoxon test confirms statistical significance, this scheme does not fully assess the model's stability across varied epidemiological periods.

3) Geographic scope: The model was validated exclusively on data from Côte d'Ivoire. Its transferability to other sub-Saharan African contexts with different climatic profiles and data quality remains to be assessed.

4) Feature set: The current model relies on a fixed set of climatic and epidemiological variables. The potential contribution of socioeconomic or behavioral features has not been explored.

F. Future Work

These limitations suggest four concrete directions. First, the correction constant should be replaced by a trend-aware mechanism such as a directional drift estimator to improve predictions during monotonic phases. Second, a walk-forward validation framework across multiple temporal windows should be adopted to rigorously assess temporal stability. Third, the approach should be evaluated across diverse geographic contexts in sub-Saharan Africa. Fourth, hybrid CBR-ML architectures that combine the interpretability of CBR with the feature-learning capacity of ML models should be explored.

VII. Conclusion

This study proposed an enhanced Case-Based Reasoning (CBR) approach for predicting malaria cases in Côte d'Ivoire, addressing two major limitations of existing models: the opacity of machine learning black-box models and the stationarity assumption of classical CBR. The proposed model, CBR_2, incorporates three key contributions : regional clustering, cyclic similarity measurement, and a temporal correction constant which jointly improve both predictive accuracy and operational interpretability.

When evaluated against seven ML models (CatBoost, XGBR, RFR, DTR, SVM, KNN, and linear regression) and two CBR baselines (CBR_0, CBR_1), CBR_2 achieves the best overall performance on the test set (RMSE = 0.081, MAE = 0.057, $R^2 = 72.40\%$), it outperformed the strongest ML competitor, RFR ($R^2 = 62.22\%$), with statistical significance confirmed by a Wilcoxon signed-rank test ($p = 1.02 \times 10^{-20}$). Beyond quantitative metrics, CBR_2 is the only model to satisfy all three qualitative criteria explainability, transparency, and adaptability making it directly deployable by health personnel without machine learning expertise. This dual advantage places CBR_2 first in the overall ranking with a score of 6, ahead of RFR (12) and CBR_1 (15).

These results confirm that it is possible to achieve competitive predictive performance while maintaining full interpretability, a critical requirement for AI adoption in public health. They also highlight the largely untapped potential of CBR in epidemiological forecasting, a field currently dominated by black-box approaches. In practice, CBR_2 can support early warning systems, epidemic anticipation, and healthcare resource allocation in resource-limited settings.

Three directions are identified for future work, improving the temporal correction mechanism, strengthening the validation framework, and extending the approach to broader geographic contexts in sub-Saharan Africa.

Acknowledgment

This work was partially supported by the international network of french-speaking engineering universities and schools RESCIF.

We would also like to thank the Ministry of Health of Côte d'Ivoire and SODEXAM for providing the data used in this study.

References

[1] E. Naroum, E. M. Maka, H. Abboubakar, P. Dayang, A. B. Bamana, B. Garga, H. D. Daouda, M. Bakouri, and I. Khan, "Comparative analysis of deep learning and machine learning techniques for forecasting new malaria cases in cameroon's adamaoua region," *Intelligence-Based Medicine*, vol. 11, p. 100220, 2025.

[2] E. Mbunge, R. C. Millham, M. N. Sibiya, and S. Takavarasha, "Application of machine learning models to predict malaria using malaria cases and environmental risk factors," in *2022 Conference on Information Communications Technology and Society (ICTAS)*. Durban, South Africa: IEEE, March 2022, pp. 1–5.

[3] M. T. Pillay *et al.*, "Utilizing a novel high-resolution malaria dataset for climate-informed predictions with a deep learning transformer model," *Scientific Reports*, vol. 13, no. 1, p. 23091, December 2023.

[4] D. Harvey, W. Valkenburg, and A. Amara, "Predicting malaria epidemics in Burkina Faso with machine learning," *PLOS ONE*, vol. 16, no. 6, p. e0253302, June 2021.

[5] M. A. Sehlabana, D. Maposa, A. Boateng, and S. Das, "Integrating climate and environmental data with Bayesian models for malaria prediction," *Statistics, Optimization and Information Computing*, vol. 14, no. 5, pp. 2930–2956, September 2025.

[6] A. K. Verma and V. Kuppili, "Data-oriented neural time series with long short-term memories (LSTM) for malaria incidence prediction in Goa, India," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Kanpur, India: IEEE, July 2019, pp. 1–6.

[7] O. Khan, J. O. Ajadi, and M. P. Hossain, "Predicting malaria outbreak in The Gambia using machine learning techniques," *PLOS ONE*, vol. 19, no. 5, p. e0299386, May 2024.

[8] A. Singh, M. Mehra, A. Kumar, M. Niranjannaik, D. Priya, and K. Gaurav, "Leveraging hybrid machine learning and data fusion for accurate mapping of malaria cases using meteorological variables in western India," *Intelligent Systems with Applications*, vol. 17, p. 200164, February 2023.

[9] D. Doyle, A. Tsymbal, and P. Cunningham, "A review of explanation and explanation in case-based reasoning," 2003.

[10] P. Pradeep, M. Caro-Martínez, and A. Wijekoon, "Empowering explainable artificial intelligence through case-based reasoning: A comprehensive exploration," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 12, pp. 7120–7139, December 2025.

[11] ——, "A practical exploration of the convergence of case-based reasoning and explainable artificial intelligence," *Expert Systems with Applications*, vol. 255, p. 124733, December 2024.

[12] M. Caro-Martínez *et al.*, "iSee: A case-based reasoning platform for the design of explanation experiences," *Knowledge-Based Systems*, vol. 302, p. 112305, October 2024.

[13] J. M. Schoenborn, R. O. Weber, D. W. Aha, J. Cassens, and K.-D. Althoff, "Explainable case-based reasoning: a survey," in *AAAI-21 workshop proceedings*, 2021.

[14] N. B. Doğan, B. U. Ayhan, G. Kazar, M. Saygili, Y. E. Ayözen, and O. B. Tokdemir, "Predicting the cost outcome of construction quality problems using case-based reasoning (CBR)," *Buildings*, vol. 12, no. 11, p. 1946, November 2022.

[15] W. Li, W. K. Härdle, and S. Lessmann, "A data-driven case-based reasoning in bankruptcy prediction," November 2022.

[16] G. A. Pérez-Pérez, M. F. Valdez-Ávila, M. G. Orozco-del Castillo, C. Bermejo-Sabbagh, and J. A. Recio-García, "CBR-FoX: A case-based reasoning software tool for auditing time series predictions," *SoftwareX*, vol. 32, p. 102450, December 2025.

[17] F. Sørmo, J. Cassens, and A. Aamodt, "Explanation in case-based reasoning–perspectives and goals," *Artificial Intelligence Review*, vol. 24, no. 2, pp. 109–143, October 2005.

[18] K. H. D. C. Monteiro *et al.*, "Integrating machine learning and spatial clustering for malaria case prediction in Brazil's Legal Amazon," *BMC Infectious Diseases*, vol. 25, no. 1, p. 802, June 2025.

[19] E. K. Panzi, N. I. Kandala, E. L. Kafinga, B. M. Tampwo, and N.-B. Kandala, "Forecasting malaria morbidity to 2036 based on geo-climatic factors in the Democratic Republic of Congo," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12271, September 2022.

[20] M. Wang *et al.*, "A novel model for malaria prediction based on ensemble algorithms," *PLOS ONE*, vol. 14, no. 12, p. e0226910, December 2019.

[21] S. W. Jalloh, B. Malenje, H. Imboga, and M. H. Hodges, "Forecasting malaria cases using climate variability in Sierra Leone," *Malaria Journal*, vol. 24, no. 1, p. 158, May 2025.

[22] J. X. Zhai *et al.*, "Development of an empirical model to predict malaria outbreaks based on monthly case reports and climate variables in Hefei, China, 1990–2011," *Acta Tropica*, vol. 178, pp. 148–154, February 2018.

[23] B. O. Nyawanda *et al.*, "Forecasting malaria dynamics based on causal relations between control interventions, climatic factors, and disease incidence in western Kenya," *Journal of Global Health*, vol. 14, p. 04208, October 2024.

[24] B. Shi, S. Lin, Q. Tan, J. Cao, and A. Et, "Inference and prediction of malaria transmission dynamics using time series data," *Infectious Diseases of Poverty*, 2020.

[25] P. Taconet *et al.*, "Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso," *Parasites & Vectors*, vol. 14, no. 1, p. 345, June 2021.

[26] S. Jinad, M. O. Kama, M. Baba-Adamu, A. C. Ikegwu, and O. Machete, "Analysis of malaria and climate in Damaturu City of Nigeria using predictive supervised learning," *Discover Sustainability*, vol. 6, no. 1, p. 1239, November 2025.

[27] M. S. Rahman and M. A. B. Shiddik, "Unraveling global malaria incidence and mortality using machine learning and artificial intelligence-driven spatial analysis," *Scientific Reports*, vol. 15, no. 1, p. 28334, August 2025.

[28] G. J. Gbaguidi, N. Topanou, W. L. Filho, and G. K. Ketoh, "Towards an intelligent malaria outbreak warning model based intelligent malaria outbreak warning in the northern part of Benin, West Africa," *BMC Public Health*, vol. 24, no. 1, p. 450, February 2024.

[29] Ç. N. Tülü and Y. Kaya, "Convolutional fourier analysis network (conv-fan-pox): A novel time–frequency approach for medical image analysis," *Biomedical Signal Processing and Control*, vol. 117, p. 109698, 2026.

[30] L. Zou, L. Xia, L. Hou, X. Zhao, and D. Yin, "Data-efficient reinforcement learning for malaria control," May 2021.

[31] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, March 2019.

[32] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications*, vol. 7, no. 1, pp. 39–59, 1994.

[33] T. W. Liao, Z. Zhang, and C. R. Mount, "Similarity measures for retrieval in case-based reasoning systems," *Applied Artificial Intelligence*, vol. 12, no. 4, pp. 267–288, 1998.

[34] A. Sylla, T. Coudert, and L. Geneste, "A case-based reasoning (CBR) approach for engineer-to-order systems performance evaluation," *IFAC-PapersOnLine*, vol. 54, no. 1, pp. 717–722, January 2021.

[35] M. F. X. Barboza *et al.*, "Prediction of malaria using deep learning models: A case study on city clusters in the state of Amazonas, Brazil, from 2003 to 2018," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 55, pp. e0420–2021, 2022.

[36] M. T. Pillay, M. T. Q. Le, Y. Takamatsu, T. V. Phong, N. Kgalane, and N. Minakawa, "Application of a Temporal Fusion Transformer and long-term climate and disease data to assess the predictive power and understand the drivers for malaria and dengue," *International Journal of Environmental Research and Public Health*, vol. 23, no. 1, p. 75, January 2026.

[37] T. C. D. Lucas *et al.*, "Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence," *Spatial and Spatio-temporal Epidemiology*, vol. 41, p. 100357, June 2022.

A Hybrid Semantic-Statistical Feature Fusion Framework for Bilingual Text Classification on Multilingual Big Data Corpora

Kavitha M[1], Dr. Purohit Shrinivasacharya[2], Dr. Y S Nijagunarya[3]
Department of Computer Science and Engineering,
Siddaganga Institute of Technology, Tumkur-572103, Karnataka, India[1,3]
Visvesvaraya Technological University, Belagavi-590018, Karnataka, India[1,2,3]
Department of Information Science and Engineering, Siddaganga Institute of Technology,
Tumkur-572103, Karnataka, India[2]

***Abstract*—As the volume of scientific information is growing exponentially in several languages, there is a need for practical and scalable bilingual classification systems for large aligned scientific text corpora. To address this challenge, this study makes two key contributions - first, a large-scale bilingual English–Hindi aligned arXiv scientific text corpus, providing a structured resource for multilingual scientific text analytics and classification research is constructed. Second, the study proposes a bilingual scientific text classification framework and performs a rigorous experimental evaluation using three strong multilingual transformer models, namely Mini Language Model-12 Layers (MiniLM-L12), Multilingual Bidirectional Encoder Representations from Transformers (mBERT), and Cross-lingual Language Model-Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach (XLM-RoBERTa), on the developed English–Hindi aligned arXiv big data corpus. English and Hindi summaries are categorized independently to investigate the performance trade-offs in each language. The proposed hybrid MiniLM-L12 + Multi-Layer Perceptron (MLP) architecture enhance the classification capability through the integration of statistical feature design with contextual sentence embeddings. Empirical analysis indicates that the proposed bilingual classification framework consistently outperforms the baseline transformer-only models, achieving a higher accuracy of 95.56% and weighted F1-score of 95.31%, while maintaining computational efficiency. The findings emphasize the effectiveness of hybrid representation learning for bilingual big data corpora and provide practical insights for scalable multilingual scholarly text analytics.**

***Keywords*—Text classification (bilingual); aligned big data corpora; multilingual transformers; mini language model-12 layers; multilingual bidirectional encoder representations from transformers; cross-lingual language model-robustly optimized bidirectional encoder representations from transformers pretraining approach; multi-layer perceptron; hybrid deep learning**

I. Introduction

Due to the continued digitization of scientific literature, the amount of scholarly literature published in different languages has increased dramatically. Although English continues to predominate in scientific communication, much of the research is reported in regional and low resource languages, especially in the present research communities. For automated information organization, retrieval, and large-scale text analytics, where conventional monolingual classification systems are ineffective, this multilingual expansion presents serious challenges. Bilingual and cross-lingual techniques for text classification, capable of operating on aligned multilingual corpora with high precision and computational efficiency are thus highly essential. Recent developments highlight transformer-based language models as the dominant methodology in multilingual natural language processing. Multilingual models like mBERT, XLM-RoBERTa, and models like MiniLM are trained to learn common semantic representations across languages and have shown strong performance on diverse multilingual NLP tasks. However, these models exhibit notable differences in terms of model size, training and inference cost, and scalability. These computational considerations are essential in large-scale text classification with a bilingual corpus, particularly in a scientific context. Furthermore, relying solely on deep transformer architectures may limit the exploitation of complementary statistical characteristics inherent in scientific texts.

Hybrid architectures offer promising alternative which can enhance efficiency and yet provide competitive classification performance by jointly using rich semantic representations together with auxiliary features using lightweight neural classifiers. Despite their potential, such approaches have received limited attention in the context of bilingual scientific text classification, particularly for aligned English-Hindi corpora.

In the current work, a comprehensive bilingual experimental study on an aligned English-Hindi arXiv scientific summary dataset is conducted. Each language is used separately and text classification experiments are carried out so that fair language-wise assessment is achieved. This study benchmark widely used multilingual transformer models, including MiniLM-L12, mBERT, and XLM-RoBERTa, and propose a hybrid MiniLM-L12 + MLP architecture that fuses contextual sentence embeddings with statistical features for enhanced performance. Experimental findings indicate that the proposed hybrid model delivers enhanced efficiency and faster inference while achieving competitive or improved accuracy and weighted F1-scores relative to larger multilingual transformer models in both languages.

The key findings of this study are summarized as follows:

- Monolingual arXiv dataset sampled to Bilingual dataset through effective translation of title and summary attributes to Hindi to obtain an aligned multi-class arXiv big data corpus.

- A comprehensive bilingual test of MiniLM-L12, mBERT, and XLM-RoBERTa on aligned English-Hindi arXiv scientific data.
- A new hybrid MiniLM-L12 + MLP model that is effective to combine contextual embeddings with statistical features in bilingual text classification.
- Analysis of language-wise performance in terms of accuracy, macro precision, macro recall, macro F1, micro precision, micro recall, micro F1, weighted precision, weighted recall and weighted F1-score measures.

The study includes valuable information regarding the effective design of bilingual text classification and demonstrates that hybrid systems establish an effective tradeoff between the performance and the computational cost of large scale multilingual scientific collections.

II. Related Work

Research on text classification has evolved through interconnected methodological streams that collectively underpin modern multilingual and large-scale classification systems. Early investigations primarily emphasized algorithmic efficiency, feature engineering, and computational scalability. Xu suggested a KNN-optimized classification scheme that minimized the use of pre-defined parameter settings [1], whereas Mao et al. proposed parameter-free compression-based technologies that mitigated the use of explicit modeling assumptions [2]. Niranjan et al. investigated structured pipelines used to categorize questions [3]. In the context of Indic multilingual text classification, Pathak and Jain introduced the μBoost framework, which integrates CatBoost with MuRIL representations to improve abusive comment detection across Indian languages. Their ensemble approach demonstrated the benefits of combining multilingual language models with structured meta-features for robust multilingual classification [4]. These early developments put in place performance standards and computation limits that fueled the transition of neural representation learning.

The development of deep neural networks was a significant breakthrough as it made it possible to extract semantic features automatically. Agbesi et al. and Malik and Jain proposed ensemble convolutional structures and knowledge-infused architectures tailored to increase robustness [5], [6]. Jamshidi et al. suggested hybrid pipelines combining transformer embeddings and sequential models [7], while Padalko et al. and Maham et al. investigated sequential and adversarial modeling for misinformation detection [8], [9]. Gao et al. and Huang et al. studied graph-based structural modeling techniques [10], [11], and Ai et al. developed semi-supervised contrastive multigraph learning [12]. Alternative neural paradigms such as Kolmogorov Arnold networks and bi-attention fusion architectures were also explored to enable flexible and adaptive classification strategies [13], [14]. Together these advances demonstrated the effectiveness of representation learning over handcrafted features, but also increased computational demands.

Transformer architectures later emerged as a dominant paradigm in text classification due to their ability to provide contextualized and generalizable representations. Joshi and Joseph analyzed efficient fine-tuning strategies [15], whereas Zhang applied multilingual multi-feature fusion architectures [16]. Bekamiri et al. presented domain-adapted transformers for patent analytics [17], and Brack et al. studied cross-domain multi-task sequential classification [18]. Jain et al. suggested hierarchical and multimodal embedding strategies on crisis datasets [19]. Nevertheless, transformer-based systems still face challenges related to computational cost, scalability, and fairness, particularly in multilingual and low-resource scenarios.

Different pretrained transformer models have become standardized architectures of cross-lingual classification in multilingual settings. Multilingual BERT (mBERT), which was trained on Wikipedia corpora involving 104 languages, using a shared set of WordPiece vocabulary, showed that transfer across languages could start being observed even without explicit alignment supervision [20]. XLM-RoBERTa (XLM-R), which was pretrained on large-scale CommonCrawl corpora in 100 languages with dynamic masked language modeling, enhanced strength and low-resource generalization with high exposure to data and training strategies [21]. Later, small distilled models like MiniLM-L12 have been created that maintain semantic expressiveness using attention and representation distillation mechanisms [22]. These models represent respectively on the efficiency-capacity axis: mBERT is able to offer consistent multilingual contextualization, XLM-R focuses on large-scale robust representation learning, and MiniLM-L12 focuses on computational efficiency and faster inference. They are appropriate reference frameworks to analyse the trade-offs in multilingual performance given their architectural variety and variations in pretraining scale.

In addition to architectural developments, multilingual studies increasingly address fairness and variability in evaluation. Cross-lingual adaptability has been explored through hybrid multilingual models [23], cross-lingual topic detection [24], and prompt-based zero-shot classification frameworks [25]. Fairness-conscious multilingual assessment has been highlighted in recent studies [26], while empirical variability in low-resource classification scenarios has been emphasized [27]. Similarly, Mehra et al. proposed PhonoBiEmbedNet, a phoneme- and bigram-based embedding framework designed for low-resource spoken word recognition. By incorporating phonetic linguistic representations and pretrained multilingual models through a late-fusion neural architecture, the approach demonstrated improved recognition accuracy under limited training data conditions [28]. These findings highlight that multilingual classification remains a complex optimization task involving linguistic diversity, data bias, and computational constraints.

The applied use of text classification includes biomedical multi-label learning [29], [30], patent analytics using hybrid transformer-graph models [17], [31], [11], and crisis informatics and misinformation detection using sequential and adversarial models [8], [9], [19]. Dataset balancing strategies for multilingual software classification have also been explored [32]. More recently, classification approaches based on large language models have introduced hierarchical taxonomy refinement and causal classification paradigms [33], [34]. Although LLM-driven systems tend to achieve strong performance, they introduce significant computational overhead and

scalability challenges in multilingual large-scale environments.

Even though there is continued improvement, there are three gaps that remain critical in the literature. To begin with, efficient bilingual frameworks that are tuned to the aligned large-scale corpora are underutilized. Second, there are not many multilingual tests with systematic fairness analysis. Third, in the real-world scientific classification environment, empirical research on efficiency-performance trade-offs between compact and large-capacity transformer architectures is under-examined. To overcome these shortcomings, scalable learning of representation should be incorporated with considerations of computational efficiency in a principled manner. By placing small distilled transformer and large-capacity multilingual models in a single bilingual assessment system, the study under consideration is expected to connect the principles of foundational scalability with the contemporary contextual representation learning and to analyze the robustness, efficiency, and fairness between English and Hindi scientific texts classification in a systematic way.

III. Proposed Methodology

This section outlines the suggested hybrid semantic-statistical feature fusion framework for bilingual scientific text classification. First, the construction and preprocessing of an aligned English-Hindi arXiv corpus to provide the semantic consistency between languages is discussed, followed by the design of a computationally efficient classification architecture that combines contextual transformer-based embeddings with statistical lexical information. Then, the mathematical description of the semantic-statistical fusion mechanism and optimisation strategy is outlined. Finally, an algorithm description of the end to end text classification workflow is presented to show how the proposed structure will be implemented.

A. Aligned Bilingual Corpus Construction and Preprocessing Strategy

The experimental evaluation of the multilingual transformer models and the proposed hybrid model is performed on the arXiv scientific corpus, a publicly accessible dataset downloadable from Kaggle: https://www.kaggle.com/datasets/sumitm004/arxiv-scientific-research-papers-dataset. The arXiv corpus is a large collection of research papers from the arXiv preprint repository, which can be used for scientific text analysis, classification, citation prediction, and trend forecasting. The dataset covers a variety of research domains, including Computer Science, Mathematics, Physics, etc. Table I provides the features of the arXiv dataset downloaded from Kaggle.

TABLE I. arXiv Dataset Features Before Preprocessing, Translation and Balancing

Dataset Feature	Description
Size	177 MB
Attributes	id, title, category, category_code, published_date, updated_date, authors, first_author, summary, summary_word_count
Modality Addressed	Text
Linguality Addressed	English
Number of Samples	136,239
Number of Categories	73

The downloaded arXiv dataset's samples featuring all attributes mentioned in Table I before preprocessing, translation and balancing is showcased in Fig. 1.

id	title	category	category_code	published_date	updated_date	authors	first_author	summary	summary_word_count
abs-2501.02725v1	artificial intelligence	artificial_intelligence	cs.AI	06-01-2025	06-01-2025	['Nanthee	'Nantheera Anan	rapid advancer	192
abs-2501.10945v1	gradient base multi	machine_learning	cs.LG	19-01-2025	19-01-2025	['Weiyu Cl	'Weiyu Chen'	multi objective	129
abs-2501.16768v1	generalization multi	machine_learning_{s	stat.ML	28-01-2025	28-01-2025	['Wen We	'Wen Wen'	multiview learn	166
abs-2501.16790v1	exponential family a	machine_learning_{s	stat.ML	28-01-2025	28-01-2025	['Kevin Ch	'Kevin Christian	self attention r	172
abs-2501.16988v2	marginal conditiona	machine_learning_{s	stat.ML	28-01-2025	29-01-2025	['Mohamr	'Mohammad Ka	interpret black	137
abs-2501.17345v1	test conditional mei	machine_learning_{s	stat.ML	28-01-2025	28-01-2025	['Yi Zhang	'Yi Zhang'	conditional me	126
abs-2501.17512v1	survey cluster base	machine_learning_{s	stat.ML	29-01-2025	29-01-2025	['Omar El-	'Omar El-Rifai'	industrial comi	163
abs-2501.17513v1	sequential learning	machine_learning_{s	stat.ML	29-01-2025	29-01-2025	['Elise CrÃ	'Elise CrÃ©pon'	study problem	122

Fig. 1. Representative samples from the original arXiv dataset displaying the metadata and textual attributes used in this study before preprocessing, machine translation into Hindi and class balancing for bilingual scientific text classification.

The dataset is preprocessed using spaCy during which title and summary text fields are cleaned by removing noise, lowercased, lemmatized, and stopwords are eliminated. Extremely short summaries are filtered and category labels are normalized and label-encoded for supervised learning. To get the aligned English-Hindi corpus, Hindi translations are obtained for title and summary attributes using the deep-translator library, which leverages Google Translate's web-based translation service. Two bilingual experts manually evaluated 1000 randomly selected translations. Approximately 90% were judged semantically equivalent to the English source text. To compare the Hindi summaries generated by Helsinki-NLP/opus-mt-en-hi model with the Hindi summaries generated by Google Translate, a Corpus BLEU score was computed. The Corpus BLEU score obtained is 0.264, which is very low, reflecting very low lexical similarity between the two translation systems, and therefore, they produce very different Hindi translations for the same English summaries. Based on the experimental observations, the Hindi summaries generated using Google Translate were selected for the subsequent experiments, as they yielded more effective classification performance.

Algorithm 1 outlines the algorithm used to create aligned English-Hindi arXiv corpus.

Algorithm 1 English–Hindi arXiv Corpus Alignment

Require: arXiv English summaries $D = \{x_1, x_2, \ldots, x_N\}$
Ensure: Aligned bilingual arXiv corpus $D^{aligned}$
Load dataset and initialize GoogleTranslator $T_{(en \to hi)}$
for each summary $x_i \in D$ **do**
  if x_i is empty **then**
    $x_i^{hi} \Leftarrow$ ""
    continue
  end if
  Translate $x_i^{hi} \Leftarrow T(x_i)$ with retry limit R
  Apply delay Δt for rate control
  Construct aligned pair (x_i, x_i^{hi})
end for
Save updated dataset as aligned bilingual arXiv corpus $D^{aligned}$

Random oversampling and augmentation are performed on this aligned English-Hindi arXiv dataset in order to obtain balanced big data corpus. Table II outlines the dataset features after the preprocessing, translation and balancing.

The preview of the arXiv dataset samples after preprocessing, translation and balancing is depicted in Fig. 2.

TABLE II. ARXIV DATASET FEATURES AFTER PREPROCESSING, TRANSLATION AND BALANCING

Dataset Feature	Description
Size	895 MB
Attributes	id, title, title_hi_ref, category, category_code, published_date, updated_date, authors, first_author, summary, summary_word_count, year, month, day, category_label, summary_hi_ref
Modality Addressed	Text
Linguality Addressed	English, Hindi
Number of Samples	296,065
Number of Categories	73

id	title	title_hi_ref	category	category_code	published_date	updated_date	authors	first_author	summary	summary_word_count	year	month	day	category_label	summary_hi_ref
abs-2205.09519v1	design math	स्वदेक ओरेगन	systems_a	eess.SY	19-05-2022	19-05-2022	['Aadhitiy	'Aadhitiya V	S neuromor	118	2022	5	19	72	न्यूरोमॉर्फिक कंप्यूटिंग में
abs-2407.18523v1	dtformer tra	डीटीफोर्मर ट्रां	machine_l	cs.LG	26-07-2024	26-07-2024	['Xi Chen',	'Xi Chen'	discrete ti	247	2024	7	26	43	असतत समय गतिशील
abs-1811.04115v1	adnet deep	एडनेट डीप नेट	multimedi	cs.MM	09-11-2018	09-11-2018	['Murhaf H	'Murhaf Hos	online vid	135	2018	11	9	50	ऑनलाइन वीडियो विज्ञा
abs-1803.10986v3	error analysi	त्रुटि विश्लेषण	numerical	cs.NA	29-03-2018	01-05-2019	['Barbara I	'Barbara Bar	popular di	191	2018	3	29	54	लोकप्रिय डीप न्यूरल नेट
abs-2212.04318v1	power consu	पावर खपत में	networkin	cs.NI	08-12-2022	08-12-2022	['Nicola Pi	'Nicola Pio	fifth gener	186	2022	12	8	51	पांचवीं पीढ़ी के रेडियो ए
abs-1809.04951v1	valid simulta	उच्च आयामी से	economet	econ.EM	13-09-2018	13-09-2018	['Philipp B	'Philipp Bac	h increase a	146	2018	9	13	28	उपलब्धता में वृद्धि उच्च
abs-1809.03343v1	distribute dy	डायनेमिक मॉड	systems_a	cs.SY	07-09-2018	07-09-2018	['Wenqing	'Wenqing Li'	large scale	198	2018	9	7	71	बड़े पैमाने पर ऑप्टिमिक
abs-2410.18605v1	understand	खिंचाड़ी की बा	machine_l	cs.LG	24-10-2024	24-10-2024	['Tianze W	'Tianze Wang	pilot studi	78	2024	10	24	43	पायलट अध्ययन एल्गोरि

Fig. 2. Representative samples from the processed multilingual arXiv dataset after preprocessing, machine translation and class balancing, highlighting the retained metadata attributes, English abstracts, translated Hindi abstracts and category labels used for bilingual classification.

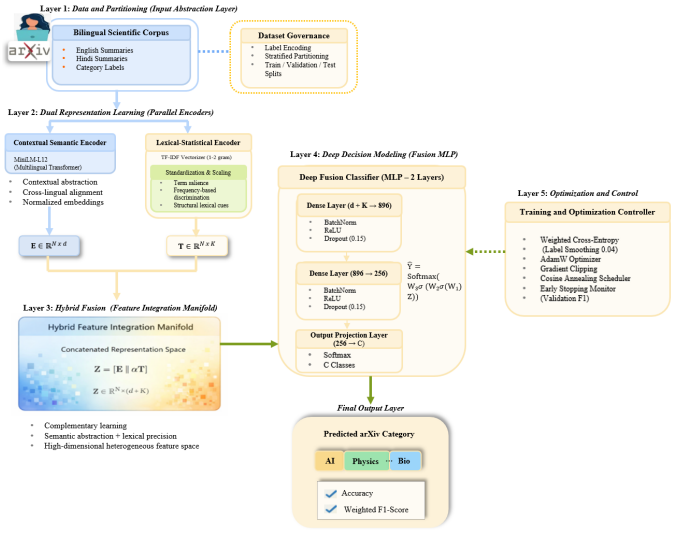


Fig. 3. Layered hybrid semantic-statistical feature fusion architecture showing multilingual context embeddings combined with statistical features, hybrid feature fusion and employing deep multiLayer perceptron to improve bilingual scientific text classification performance.

This aligned English-Hindi arXiv corpus is split using stratified sampling into 64% training, 16% validation, and 20% test sets to preserve the class distribution across all partitions. Experimental evaluation on this aligned big data corpus is performed separately for each language using multilingual transformer models and the proposed hybrid model to assess the language-specific performance.

B. Architectural Overview

To effectively address multilingual scientific summary classification, this study proposes a hybrid semantic-statistical fusion architecture that integrates contextual transformer-based representations with complementary lexical features within a unified deep learning framework. The proposed architecture, in contrast to purely embedding-driven models, further increases the discriminative power by explicitly modeling semantic abstraction as well as the statistical relevance of terms. This structure is specifically useful in scientific abstracts, where a mismatch in class representation, the variability of structure and vocabulary are significant challenges.

The proposed model depicted in Fig. 3 presents a hybrid representation learning model that works well to classify scientific literature by considering both deep contextual semantics and statistical lexical evidence. This system starts with a bilingual scientific corpus, i.e. English titles and summaries with category labels. A governance process of the data is necessary in order to guarantee consistency and reproducibility of the experimental pipeline whereby categorical labels are coded into numerical values and the dataset is stratified with stratified sampling. This stratification maintains the underlying distribution of classes in the training, validation, and test subsets. Hence, it ensures sound performance measurement and limits the chances of sampling bias in the process of training a model.

After the preparation of data, the architecture uses a dual representation learning approach where two complementary encoders are used simultaneously to represent various aspects of a textual information. The former route involves the use of a multilingual MiniLM-L12 transformer encoder, which produces contextualized representations of each document in the form of embeddings. This encoder represents deep semantic connections between tokens through self-attention mechanisms to enable the representation to absorb contextual dependency, semantic abstraction, and cross-lingual alignment that may exist within scientific text. The resulting dense embedding captures the overall semantic structure of the document, allowing the model to identify conceptual similarities despite the use of different lexical expressions.

Simultaneously, a lexical-statistical encoder is built to create a complementary representation with TF-IDF vectorization. The component uses discriminative lexical signals to extract the terms according to their significance in relation to their frequency in the corpus. In this way, domain-specific terminology which strongly defines specific scientific domains is emphasized more. TF-IDF, in contrast to contextual embeddings, which learn meaning by interacting with the context, emphasizes statistically salient keywords and structural lexical patterns that are often good predictors of types of documents. With this lexical view, the architecture conserves important information that would otherwise be diluted in neural embeddings.

The transformer encoder produces contextual semantic embeddings, while the TF-IDF module generates lexical-statistical feature vectors that encode term importance and document-specific vocabulary patterns. These complementary representations are integrated through a hybrid feature fusion layer utilizing feature concatenation, thereby constructing a unified high dimensional representation space. In particular, the semantic embedding vector and lexical-statistical feature vector are concatenated along the feature dimension, maintaining both the contextual semantic knowledge and discriminative lexical evidence. The resulting fused representation preserves

heterogeneous information sources that jointly characterize the conceptual content and terminological specificity of scientific documents. Such integration is particularly useful for scientific document classification, where semantic understanding and domain specific vocabulary play important roles in accurate category discrimination.

The fused representation is thereafter fed into a deep decision modeling component that is a multi-layer perception-based classifier. This type of fusion classifier also learns non-linear interactions among the semantic and lexical representations that allow the model to detect complex patterns that might not be discovered when these heterogeneous cues are used individually. The classifier gradually optimizes the fused representation to a discriminative feature space that is appropriate to predict categories by successive dense transformations, with the aid of the batch normalization and rectified linear activation functions. The dropout regularization is used to reduce the effect of overfitting by promoting feature learning that is robust throughout the learning process. The classifier effectively fills the gap between neural semantic representations and traditional lexical indicators by modeling interactions between contextual and statistical signals.

The architecture is coupled with a specific optimization and training control mechanism that is in charge of the learning dynamics of the model. This controller controls updating of parameters by adaptive optimization and training stabilization by gradient management and training scheduling strategies. Weighted cross-entropy is used to deal with possible class imbalance in the scientific data, so that no minority categories have a disproportionate influence on the learning goal. Gradient clipping ensures that deep networks do not update unstably, whereas cosine-based learning rate scheduling brings about a slow adaptation of optimization dynamics that encourages smooth convergence. The early stopping methodology which relies on F1-score of validation further ensures that the training process stops as soon as the generalization performance stops improving and thus avoiding overfitting and maintaining the robustness of the model.

Lastly, the trained classifier is used to generate the inferred scientific category of each document by mapping the learned representation into the target label space. The accuracy and weighted F1-score are used to measure the performance of the model which can give a holistic analysis of the effectiveness of classifications under imbalanced categories. Through the integration of transformer-based semantic modeling with statistical lexical representations and their combination by the means of a deep fusion-classifier, the given architecture creates a powerful framework of the scientific document categorization that exploits the strengths of both neural and traditional text representation paradigms.

C. Mathematical Formulation of the Proposed Semantic-Statistical Fusion Model

The suggested classification framework combines contextual semantics representations and statistical lexical representations to achieve a hybrid feature space of scientific text classification. Although transformer-based encoders can provide deep contextual semantics, they might not be able to maintain the significance of domain-specific lexical clues which can be very important in scientific texts. The proposed model overcomes this weakness by integrating multilingual MiniLM semantic embeddings with normalized TF-IDF statistical representations and learns their interactions by a simple FusionMLP classifier. This composite formulation allows the model to take advantage of both contextual meaning and term-level significance in a unit representation space. The following mathematical framework formally describes the construction of the semantic-statistical representation, the feature fusion mechanism, the classification function, and the optimization strategy used to learn the proposed model.

Let x_i denote the i-th document in the dataset. The proposed framework constructs a hybrid representation by combining contextual semantic embeddings obtained from a multilingual MiniLM encoder with normalized statistical TF–IDF features.

1) Semantic representation: Each document is first encoded using the MiniLM transformer to obtain contextual semantic embeddings:

$$e_i = f_{\text{MiniLM}}(x_i)$$

where, $f_{\text{MiniLM}}(\cdot)$ denotes the pretrained MiniLM encoder and $e_i \in \mathbb{R}^d$ represents the semantic embedding of dimension d.

2) Statistical feature extraction: To capture lexical importance signals, TF–IDF features are extracted for each document:

$$t_i = \text{TFIDF}(x_i), \qquad t_i \in \mathbb{R}^K$$

The statistical features are standardized using training-set statistics:

$$\tilde{t}_i = \frac{t_i - \mu}{\sigma}$$

where, $\mu = \mathbb{E}[t_{\text{train}}]$ and $\sigma = \text{Std}[t_{\text{train}}]$ denote the mean and standard deviation computed from the training corpus.

3) Controlled statistical feature scaling: To regulate the contribution of statistical features relative to semantic embeddings, a scaling coefficient is applied:

$$s_i = \alpha \tilde{t}_i \tag{1}$$

where, $\alpha \in [0, 1]$ controls the influence of the statistical component.

This transformation ensures that statistical features remain normalized and proportionally integrated with semantic embeddings.

4) Hybrid semantic-statistical feature fusion: The semantic and statistical representations are concatenated to construct the hybrid feature vector:

$$z_i = F(x_i) = [e_i \,;\, s_i] \tag{2}$$

where, $z_i \in \mathbb{R}^{d+K}$ represents the fused feature vector integrating contextual semantic information and lexical statistical signals. Using Eq. (1), the hybrid representation combines semantic embeddings with scaled statistical features.

5) FusionMLP decision function (two-layer architecture): The fused representation is passed through a two-layer multilayer perceptron (MLP) classifier to capture interactions between semantic and statistical components:

$$h_1 = \sigma(W_1 z_i + b_1)$$

$$h_2 = \sigma(W_2 h_1 + b_2)$$

$$\hat{y}_i = \text{Softmax}(W_3 h_2 + b_3)$$

where, z_i is the fused semantic–statistical feature vector defined in Eq. (2); W_1, W_2, W_3 and b_1, b_2, b_3 denote the learnable weight matrices and bias vectors of the MLP layers; $\sigma(\cdot)$ represents a nonlinear activation function (e.g., ReLU); h_1 and h_2 correspond to intermediate hidden representations; and $\hat{y}_i$ denotes the predicted class probability distribution obtained through the Softmax layer which can be compactly expressed as

$$\hat{y}_i = g_\theta(F(x_i))$$

where, $\theta = \{W_1, W_2, W_3, b_1, b_2, b_3\}$ denotes the model parameters.

6) Hybrid representation learning objective: The model parameters are optimized using cross-entropy loss as defined in Eq. (3):

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} y_i \log g_\theta(F(x_i)) \tag{3}$$

where, N denotes the total number of training samples, y_i represents the ground-truth label of the i-th sample, $g_\theta(\cdot)$ denotes the classifier parameterized by θ, $F(x_i)$ is the fused semantic–statistical representation defined in Eq. (2), and $\mathcal{L}(\theta)$ represents the cross-entropy loss used to optimize the model parameters.

This objective encourages the classifier to learn discriminative decision boundaries within the hybrid feature space defined in Eq. (2).

7) Optimization strategy: Training is performed using cosine annealing with warm restarts for adaptive learning-rate scheduling which is expressed in Eq. (4):

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{\pi T_{\text{cur}}}{T_i}\right)\right) \tag{4}$$

where, η_t denotes the learning rate at iteration t, $\eta_{\min}$ and $\eta_{\max}$ represent the minimum and maximum learning rates respectively, T_{cur} denotes the number of iterations since the last restart, and T_i denotes the length of the i-th cosine annealing restart cycle, representing the number of iterations over which the learning rate decays from $\eta_{\max}$ to $\eta_{\min}$ before a restart occurs, computed as

$$T_i = T_0 \cdot T_{\text{mult}}^k$$

here, T_0 denotes the initial restart period, T_{mult} is the restart cycle multiplier controlling the growth of subsequent cycles, and k indicates the number of completed restart cycles.

8) Optimal model selection criterion: The final classifier is selected by maximizing validation performance which is defined in Eq. (5):

$$M^* = \arg\max_M F1_{\text{val}}(M) \tag{5}$$

where, $F1_{\text{val}}$ denotes the validation F1-score.

9) Compact unified formulation of the proposed model: The complete semantic–statistical fusion learning framework can be summarized as in Eq. (6):

$$M^* = \arg\max_M F1_{\text{val}}\left(g_\theta\left[f_{\text{MiniLM}}(x_i) \,;\, \alpha\left(\frac{\text{TFIDF}(x_i) - \mu}{\sigma}\right)\right]\right) \tag{6}$$

This formulation integrates multilingual contextual embeddings and normalized statistical features into a unified hybrid representation for efficient bilingual scientific document classification. The suggested model has three methodological contributions:

- A structured semantic-statistical fusion formulation that combines contextual embedding with features of lexical importance into a shared representation space.
- A controlled statistical scaling mechanism that regulates the influence of TF-IDF features to maintain distributional balance with semantic embeddings.
- An efficient hybrid representation learning strategy, which captures nonlinear interactions between semantic and statistical signals using a lightweight neural classifier.

D. Algorithm for Text Classification

The overall workflow of the proposed semantic–statistical feature fusion framework is summarized in Algorithm 2. It outlines the end-to-end pipeline for aligned bilingual text classification, including data preprocessing, hybrid feature construction using MiniLM embeddings and TF–IDF statistics, classifier training, and final evaluation.

Algorithm 2 Semantic–Statistical Feature Fusion Framework for Aligned Bilingual Text Classification

Require: Aligned document summaries $D = \{x_1, x_2, \ldots, x_N\}$
Ensure: Optimal classifier M^* and evaluation metrics $\{\text{Accuracy}, \text{Weighted F1-score}\}$
Load dataset D and corresponding class labels $Y = \{y_1, y_2, \ldots, y_N\}$
Encode labels and perform stratified partitioning to obtain training, validation, and test sets
for each document $x_i \in D$ **do**
    Generate contextual semantic embedding e_i using the MiniLM encoder
    Compute TF–IDF vector t_i using unigram and bigram statistics and normalize using training-set statistics
    Construct fused representation $f_i = [e_i, t_i]$
end for
Feed fused features f_i into the FusionMLP classifier M
Train M using cross-entropy loss with AdamW optimization and cosine learning-rate scheduling
Monitor validation F1-score and apply early stopping to obtain optimal model M^*
Evaluate M^* on the test set and compute Accuracy and Weighted F1-score

IV. Empirical Evaluation and Comparative Performance Analysis

A comprehensive empirical study of the suggested semantic-statistical feature fusion model on English-Hindi aligned corpus of arXiv is presented in this section. Model specification and hyperparameter configurations are outlined in detail. The quantitative performance of the proposed model, its' convergence behavior, class-level robustness, overall classification performance, architectural contribution via ablation, computational efficiency, and bilingual alignment properties are systematically analyzed. Multiple evaluation metrics are used to report comparative measurements against known multilingual transformer baselines in order to provide a rigorous and balanced performance evaluation.

A. Model Specification and Training Parameters

Using the common data splits and hyperparameter configuration, both the baseline and proposed models were trained to carry out comprehensive testing of the proposed framework. In this subsection, the architectural considerations of the base and hybrid models are presented as well as the training environments such as the learning rates, the batch size and the optimization methods. The setup allows studying convergence processes and performance efficiency and ensures that the proposed MiniLM-L12+MLP fusion architecture and the multilingual baselines can be fairly compared. Table III summarizes the entire experimental arrangement used for classification.

TABLE III. Model Design, Training Protocol and Hyperparameter Configuration for Hybrid Feature Fusion Classification Model Applied to an Aligned English–Hindi arXiv Corpus

Category	Hyperparameter	Value	Purpose
Embedding Model	Sentence Transformer	paraphrase-multilingual-MiniLM-L12-v2	Lightweight multilingual semantic encoder.
	Embedding Normalization	Enabled	Ensures cosine-consistent representations.
Batching	Batch Size	96	Balances GPU memory & convergence stability.
Training	Epochs (max)	65	Allows full convergence with early stopping.
	Learning Rate	1.5×10^{-4}	Stable for fused semantic + lexical features.
	Optimizer	AdamW	Handles sparse & dense feature fusion.
	Weight Decay	3×10^{-5}	Prevents overfitting.
	Gradient Clipping	1.0	Stabilizes training.
Regularization	Dropout	0.15	Controls co-adaptation in fusion layers.
	Label Smoothing	0.04	Improves generalization.
Scheduler	LR Scheduler	CosineAnnealing Warm Restarts	Escapes sharp minima.
	Initial Restart Period (T_0)	8	Fast warm restarts.
	Restart Multiplier (T_{mult})	2	Gradually longer cycles.
Early Stopping	Patience	8 epochs	Prevents overfitting.
	Monitoring Metric	Weighted F1-score	Robust for class imbalance.
Classifier	Architecture	MLP ($896 \rightarrow 256 \rightarrow C$)	Deep fusion classifier.
	Batch Normalization	Yes	Training stability.
Evaluation	Metrics	Accuracy, Weighted F1-score	Balanced performance reporting.
	Hardware	GPU	Efficient computation.

The multilingual MiniLM-L12 encoder computes a semantic embedding vector of 384 dimensions for each document in English text classification. Meanwhile, a 512-dimensional TF-IDF representation with the help of unigram and bigram features is obtained after removing English stop-words. Lexical information is also preserved by scaling the TF-IDF features by a factor of 0.50, while semantic embeddings dominate the features. The output of the semantic embedding vector is concatenated with the output of the scaled TF-IDF feature vector, forming an 896-dimensional representation (384 + 512), and then input to the residual MLP classifier.

The multilingual MiniLM-L12 encoder also generates a 384-dimensional vector embedding of the text in Hindi for the task of Hindi text classification. No stop-word filtering is done to retain the linguistic cues for Hindi and a 512 dimensional unigram–bigram TF-IDF representation is extracted. After

obtaining the semantic embedding vector, the features in the TF-IDF are scaled by a factor of 0.50 and then fused at the feature level. The fused representation from this process is a 896 dimensional (384+512) vector, fed into the residual MLP classifier.

B. Quantitative Performance Evaluation of the Proposed Framework

To evaluate the efficiency of the proposed semantic-statistical hybrid structure, quantitative test of English and Hindi summary classification is done which is outlined in this section. Performance is evaluated using standard classification measures derived from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). For balanced assessment of the potentially imbalanced scientific categories, Accuracy and Weighted F1-score are used.

Classification accuracy is computed as in Eq. (7):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

For class-wise evaluation, precision, recall, and F1-score are defined as in Eq. (8), (9) and (10) respectively:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{8}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{9}$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{10}$$

To account for class imbalance, the weighted F1-score is calculated as in Eq. (11):

$$\text{Weighted-F1} = \sum_{c=1}^{C} \frac{n_c}{N} \cdot \text{F1}_c \tag{11}$$

where, C denotes the total number of classes, n_c represents the number of samples in class c, and N denotes the total number of samples.

The findings in Table IV indicate the general categorization ability of the proposed MiniLM-L12 + MLP (2 Layers) + TF-IDF fusion model in the bilingual environment. The proposed framework delivers high and consistent results between the two languages, reaching an accuracy of 95.56% and a weighted F1-score of 95.31% in English, and 94.85% in accuracy with a weighted F1-score of 94.53% in Hindi. The cost of marginal difference in performance between English and Hindi represents the cross-lingual generalization and strong semantic correspondence. The high weighted F1-scores also prove that the model has balanced predictive accuracy across various scientific categories indicating the success of the semantic-statistical feature fusion strategy in the classification of bilingual scientific texts.

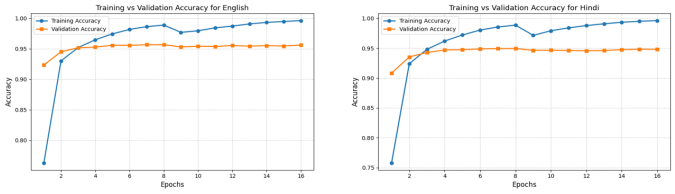
TABLE IV. ACCURACY AND WEIGHTED F1-SCORE ANALYSIS OF THE PROPOSED MODEL ON ENGLISH AND HINDI SUMMARY CLASSIFICATION

	Accuracy		Weighted F1-Score	
Proposed Model	**English**	**Hindi**	**English**	**Hindi**
MiniLM-L12 + MLP (2 Layers) + TF-IDF	0.9556	0.9485	0.9531	0.9453

C. Training Dynamics and Convergence Behaviour of the Proposed Model

This subsection analyses the learning stability and convergence characteristics of the proposed bilingual MiniLM-L12 + MLP framework using Training Accuracy curves, Validation F1 curves, and Training-Validation Loss curves across epochs. All these metrics give a clue of the efficiency in optimization, generalization, and overfitting behavior of both English and Hindi datasets.

The training dynamics are characterized by a quick early convergence and a slow stabilization in subsequent epochs, which is evidence of efficient parameter adaptation and stable optimization. Fig. 4(a) and 4(b) shows that the training accuracy is steep in the initial few epochs, which is indicative of learning features effectively and using a strong gradient-driven optimization. Following the initial phase, the curve switches to smooth plateau towards almost saturation (∼99%), which proves to be stable convergence without oscillations. The fact that the two languages are consistent demonstrates the strength of the bilingual representation learning.

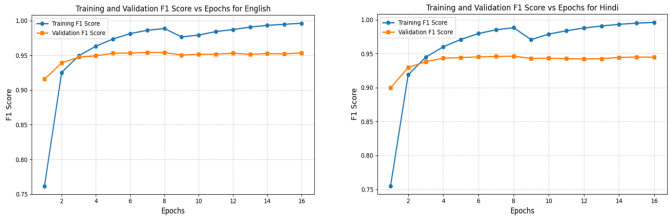


(a) For English Language (b) For Hindi Language

Fig. 4. Accuracy of the proposed model with progression of training epochs, which showed rapid convergence, low overfitting, and stable generalization in English and Hindi languages.

The validation F1 score increases consistently through the initial epochs and reaches the high level (∼95%), meaning that the model has a high-performance in terms of generalization as discussed in Fig. 5(a) and 5(b). The narrow discrepancy between training and validation curves indicates regulated model capability and restricted overfitting. The trend of the validation F1 also confirms that the classifier has a balanced precision recall trade-offs across classes.

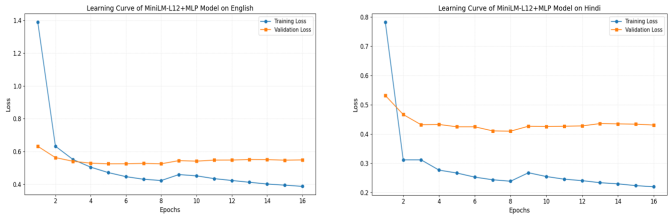
Fig. 6(a) and 6(b) show the loss-versus-epochs curves of both English and Hindi datasets, which approach zero very quickly in early epochs, which is evidence of an effective gradient based learning and fast parameter adjustment. After this initial convergence stage, loss curves begin to flatten which is an indication of a stabilization in the process of optimization. Both the validation loss and the training loss are stable and have slight variation indicating model capacity is controlled and regularization is successful. The lack of a substantial sep-



(a) For English Language (b) For Hindi Language

Fig. 5. Validation F1-score of the proposed model with training epochs, showing the rapid start with a high convergence, followed by a level with stabilization, and then high performance in generalization in both English and Hindi language.

aration between training and validation paths is a verification of the low overfitting and high-quality generalization both in languages.



(a) For English Language (b) For Hindi Language

Fig. 6. Training and validation loss traces per epoch of the proposed model in English and Hindi summary classification, showing early rapid convergence and stable optimization behavior.

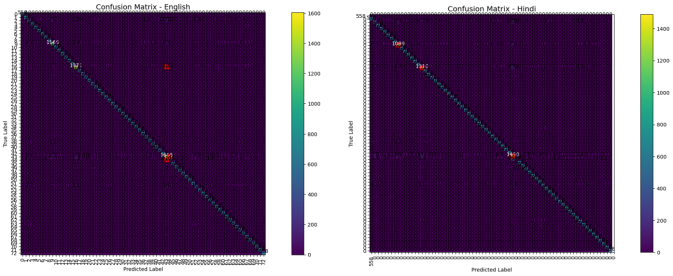
In general, the dynamics of the losses observed support the stability and robustness of the proposed framework. The fast convergence and a low validation loss show that the model does not only learn discriminative features effectively but also retains the performance of cross-lingual generalization. Such a consistent optimization pattern is also additional evidence supporting the power of the MiniLM-L12 + MLP model in the task of bilingual scientific text classification.

D. Class-Level Performance Analysis of the Proposed Model

In this subsection, a comprehensive analysis of the proposed MiniLM-L12 + MLP model on a per-class level is conducted to determine its discriminative ability when dealing with individual categories. Although overall accuracy and macro-F1 give a comprehensive picture of the performance, a more detailed analysis of the results by classes gives a better understanding of the label-specific behavior, the patterns of inter-class confusion, and the cross-lingual consistency.

Here, confusion matrices of English and Hindi are studied and per-class F1-scores are investigated and a comparative visual representation of English and Hindi class-wise F1 performance is given. These analyses collectively demonstrate that the model ensures a balanced precision-recall trade-offs between categories, detects the possible confusion-prone classes, and confirms the strength of the bilingual representation learning framework.

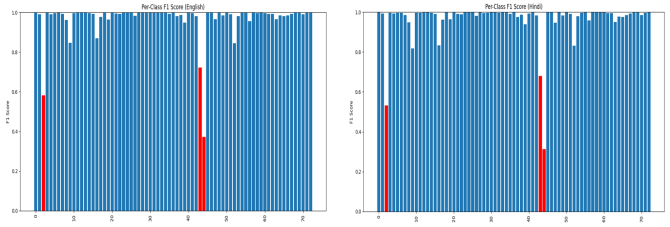
The confusion matrices provided in Fig. 7(a) and 7(b) of both English and Hindi reveal high degree of concentration



(a) For English Language (b) For Hindi Language

Fig. 7. Confusion matrix analysis of the proposed model for English and Hindi text classification, illustrating the distribution of true and predicted labels, highlighting strong category discrimination with reduced inter-class confusion.

of all the values along the diagonal which implies that a majority of the instances are being correctly classified. The low values of off-diagonals indicate low levels of misclassification of classes. The pattern indicates high accuracy, and the performance of the model in the two languages and also shows that multilingual text classification is well learned and capable of good generalization.



(a) For English Language (b) For Hindi Language

Fig. 8. Per-class F1 score analysis of the proposed model for English and Hindi text classification, demonstrating balanced classification performance, confirming the effectiveness of hybrid feature representation.

In order to further justify this performance at a more detailed level, the per-class F1 scores are examined. Although the confusion matrices give a general picture of the correct and incorrect predictions, the per-class F1 score gives a clear picture of the ability of the model to strike the right balance between precision and recall on each category selected. Based on the Fig. 8(a) and 8(b), English and Hindi models have consistently high F1 scores on most of the classes, which shows that the models have high and consistent classification performance. There are only a few classes with slightly lower values, indicating that there is a slight room to improve. In general, the findings indicate that the model is well generalized in both languages, and the performance is strong across categories.

After analyzing the individual per-class version of the F1 score, the comparative graph in Fig. 9 gives a direct class-based analysis of English and Hindi. According to the visualization, the proposed model demonstrates the same high F1 scores on the majority of the classes in both languages, with the performance trends being slightly different. Small differences identified in some classes suggest some slight language-specific differences, but in general, the outcomes prove that the model is highly multilingual and has equal

classification power when working in both English and Hindi datasets.

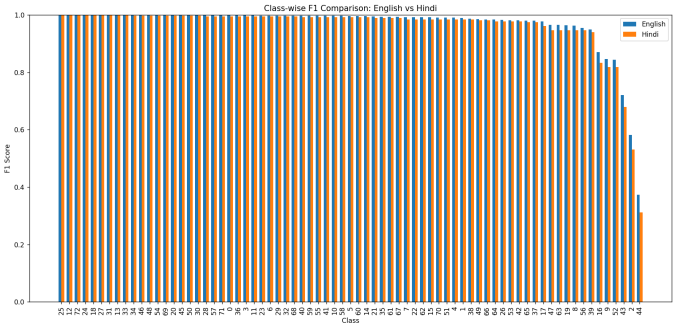


Fig. 9. Class-wise F1 score comparison of the proposed model for English and Hindi arXiv summary classification, highlighting strong cross-lingual generalization.

E. Overall Comparative Performance Analysis with Baselines

This subsection presents a comprehensive quantitative comparison between the proposed MiniLM-L12 + MLP fusion framework and established multilingual transformer baselines (MiniLM-L12, mBERT, and XLM-RoBERTa) on the English-Hindi aligned arXiv corpus. In order to ensure a balanced assessment across class distributions and languages, performance is evaluated using accuracy, macro, micro, and weighted precision, recall, and F1-scores The comparison will focus both on the absolute performance increase and also on the consistency and strength of the proposed approach in bilingual environments.

The macro F1-score, which treats all classes equally, is computed as in Eq. (12):

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{12}$$

where, C denotes the total number of classes, and $F1_c$ represents the F1-score of class c.

The micro F1-score, which aggregates global counts across classes, is given by Eq. (13):

$$\text{Micro-F1} = \frac{2\sum_c TP_c}{2\sum_c TP_c + \sum_c FP_c + \sum_c FN_c} \tag{13}$$

As demonstrated in Table V, the suggested MiniLM-L12 + MLP architecture yields the best performance in almost all the evaluation metrics of both the English and Hindi languages. The macro and micro F1-scores also validate the stability of performance in categories illustrating that there is no dominance in the improvement but it is always distributed throughout the spectrum of classes. The improvements to the strongest baseline (XLM-RoBERTa) are also consistent across all languages and all aggregated measures, indicating that semantic-statistical feature fusion can be successfully used to reinforce multilingual scientific text classification. In general, the findings indicate excellent, balanced, and statistically strong bilingual performance of the suggested framework in comparison to the state-of-the-art transformer baselines.

TABLE V. COMPARATIVE PERFORMANCE ANALYSIS OF BASELINE TRANSFORMER MODELS AND THE PROPOSED MINILM-L12+MLP FUSION FRAMEWORK FOR ENGLISH–HINDI ALIGNED ARXIV CORPUS

Performance	Model							
	MiniLM		mBERT		XLM-R		Proposed MiniLM+MLP	
Metric	E	H	E	H	E	H	E	H
Accuracy	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Macro Prec.	0.9245	0.9058	0.9561	0.9476	0.9530	0.9480	**0.9670**	**0.9603**
Macro Recall	0.9413	0.9242	0.9703	0.9659	0.9684	0.9661	**0.9714**	**0.9678**
Macro F1	0.9305	0.9122	0.9623	0.9554	0.9594	0.9558	**0.9684**	**0.9634**
Micro Prec.	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Micro Recall	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Micro F1	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Weighted Prec.	0.9087	0.8872	0.9443	0.9351	0.9417	0.9361	**0.9522**	**0.9437**
Weighted Recall	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Weighted F1	0.9109	0.8899	0.9424	0.9329	0.9389	0.9337	**0.9531**	**0.9453**

To further substantiate the comparative findings reported in Table V, a percentage gain analysis is conducted to quantify the relative improvement of the proposed MiniLM-L12 + MLP framework over established multilingual baselines. Although absolute performance measures reveal unwavering superiority in the English and Hindi tasks, relative gain calculation offers some normalized view of the magnitude of improvement in models with different baseline levels.

The percentage gain is computed as in Eq. (14)

$$\text{Gain}(\%) = \frac{M_{\text{proposed}} - M_{\text{baseline}}}{M_{\text{baseline}}} \times 100 \tag{14}$$

where, M denotes the corresponding evaluation metric. This formulation ensures that improvements are measured relative to the baseline strength rather than in absolute terms. These are the percentage-based improvements in both accuracy and weighted F1-score, which are represented in the heatmap in Fig. 10, which gives a more accurate view of differences in performance between the proposed framework and MiniLM-L12, mBERT, and XLM-RoBERTa.

As indicated in the heatmap, the suggested framework has the highest relative gains over MiniLM-L12, with the improvements of 4.06% (English accuracy), 5.59% (Hindi accuracy), 4.63% (English F1), and 6.23% (Hindi F1). Such large gains signify that TF-IDF-based statistical features combined with contextual embeddings are powerful boosters of representational capabilities of small transformer architecture. The improvement when compared with mBERT is more moderate, and it is about 1.13-1.33 % showing the growth in form of improvements to already strong multilingual benchmark. On the XLM-RoBERTa, the strongest baseline, the improvements are, nevertheless, positive, ranging between 1.15 and 1.51%, which proves the idea that large-capacity pretrained models can still be influenced by semantic-statistical fusion.

Significantly, not only accuracy, but also weighted F1-score is being shown to be continuously improved. The consistent improvement in both languages and both measures proves that the given framework improves the overall predictive ability and its strength even in the case of uneven distributions of

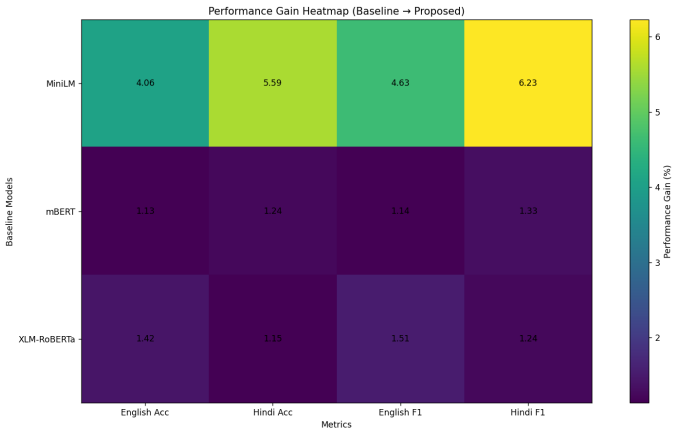


Fig. 10. Heatmap of percentage performance gains of the proposed model over MiniLM-L12, mBERT, and XLM-RoBERTa, highlighting consistent improvements across English and Hindi classification accuracy and weighted F1-scores.

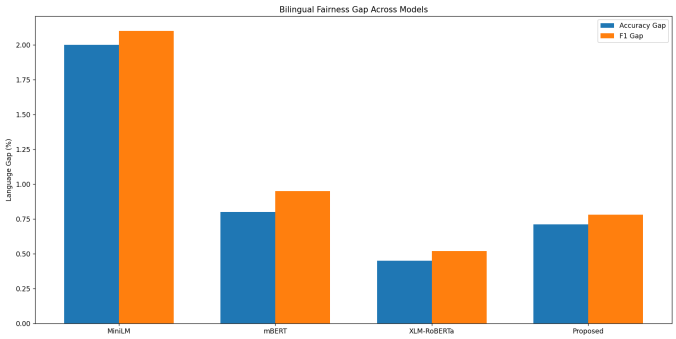


Fig. 11. Bilingual fairness gap comparison that demonstrates the performance disparity between English and Hindi languages in baseline and proposed models.

scientific categories. Taken together, the heatmap confirms the excellence and consistency of the proposed bilingual fusion model at different transformer capacities.

F. Cross-Language Alignment and Fairness Results in Comparison with Baseline and Proposed Models

In order to measure the strength and impartiality of the suggested scheme of bilingual classification, the cross-language alignment of English and Hindi is studied through performance gap analysis. Instead of measuring absolute performance, this subsection measures the consistency of individual model performance across languages. A lower performance difference shows a more robust cross-lingual representation balance and enhanced fairness, which ensures that a gain in one language does not affect the other in a more inappropriate way. This assessment is especially significant in multilingual scientific text analytics, where balanced cross-language performance is a direct measure of representation quality and transferability.

The bilingual fairness gap is computed as the absolute difference between English and Hindi performance metrics which is as defined in Eq. (15):

$$\text{Gap}_{metric} = |M_{EN} - M_{HI}| \tag{15}$$

where, M_{EN} and M_{HI} denote the evaluation metric computed for the English and Hindi text respectively, and M can represent either Accuracy or F1-score. The absolute operator ensures non-negativity and measures deviation irrespective of direction.

Based on the results obtained as depicted in Fig. 11, the proposed MiniLM-L12 + MLP fusion model proves to have a significantly small bilingual gap and, at the same time, achieves the highest overall Accuracy and F1 scores. This is a good sign of cross language embedding compatibility and good features fusion. The proposed architecture does not only provide competitive multilingual consistency when compared to bigger models like mBERT and xLM-RoBERTa, but also enhances efficiency. The decreased fairness disparity puts to rest the fact that the lightweight hybrid modeling can gain balanced and fair bilingual performance without declining predictive strength, which lays down the suitability of lightweight hybrid modeling as an approach to scale-oriented multilingual scientific document classification.

According to the bilingual alignment scatter plot in Fig. 12, it is evident that all models have high-levels of cross-lingual consistency because all their coordinate pairs are consistent. $(x_i, y_i) = (\text{Accuracy}_{EN}, \text{Accuracy}_{HI})$ lie close to the ideal alignment line $y = x$. The deviation from perfect alignment is quantified as $\Delta_i = |x_i - y_i|$, where smaller Δ_i indicates better bilingual stability.

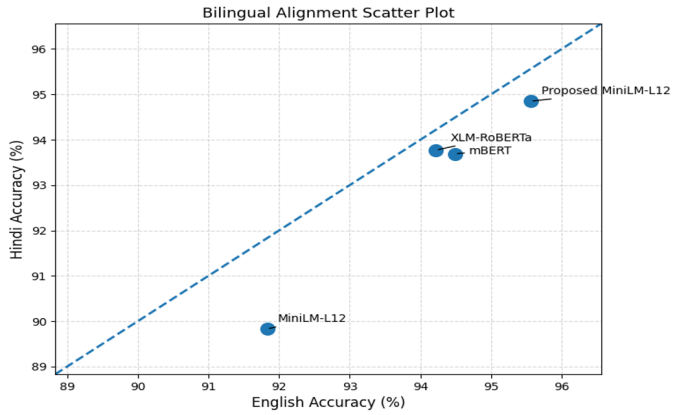


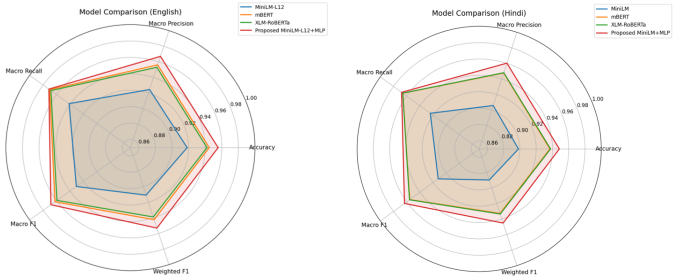
Fig. 12. Bilingual alignment scatter plot comparing the English and Hindi classification accuracy of baseline models and the proposed framework with the diagonal line representing an ideal cross-language performance balance.

The proposed MiniLM-L12 + MLP model is positioned closest to the top-right region, meaning it maximizes $\max(x_i, y_i)$ simultaneously while maintaining minimal alignment gap, demonstrating strong balanced performance across English and Hindi. Regarding the baselines, MiniLM-L12 is one of the weakly distilled transformer baselines with fast inference and relatively lower accuracy; mBERT and XLM-RoBERTa are powerful multilingual transformer baselines that demonstrate high accuracies because of the deep contextual representations; however, their points, although having a tendency to be close to the diagonal, have more significant differences in the alignment and do not exceed the suggested model in terms of combined bilingual capabilities. It means

that despite such large pretrained multilingual models being strong, the suggested hybrid structure demonstrates an even better cross lingual efficiency accuracy ratio and much lower processing cost.

The radar plots depicted in Fig. 13 provide a holistic multi-metric comparison by representing each model as a vector in a 5-dimensional performance space:

$$M_i = [Acc,\ P_{\text{macro}},\ R_{\text{macro}},\ F1_{\text{macro}},\ F1_{\text{weighted}}]$$



(a) For English Language (b) For Hindi Language

Fig. 13. Dual-language radar chart comparing MiniLM-L12, mBERT, XLM-RoBERTa, and the proposed MiniLM-L12+MLP model across English and Hindi summary classification accuracy, macro precision, macro recall, macro F1 and weighted F1-scores, illustrating the superior and balanced bilingual performance achieved by the proposed framework.

Each metric is mapped to a polar coordinate axis with equally spaced angles as defined in Eq. (16):

$$\theta_k = \frac{2\pi k}{K}, \quad k = 0, 1, \ldots, K-1 \tag{16}$$

where, θ_k denotes the angular position of the k-th metric on the circular axis, with $K = 5$ evaluation metrics. The radial value r_{ik} corresponds to the metric score of model i for metric k. where $K = 5$ metrics. The radial value r_{ik} corresponds to the metric score of model i for metric k. By closing the polygon (repeating the first value), the area enclosed by each model can be interpreted as an approximate aggregate performance measure, defined in Eq. (17):

$$\text{Area}_i \propto \sum_{k=1}^{K} r_{ik}\, r_{i(k+1)} \sin\left(\frac{2\pi}{K}\right) \tag{17}$$

A larger enclosed area indicates stronger overall balanced performance across metrics. In both English and Hindi plots depicted in Fig. 13, the proposed MiniLM-L12 + MLP constantly optimizes the outermost and most homogenous polygon implying concurrent maximization of accuracy, precision, recall, and F1 values. All the baselines are also characterized by high and stationary performance where, mBERT and XLM-RoBERTa have competitive macro-level results which are provided by deep multilingual representations. Nevertheless, it is possible to state that the offered hybrid model will provide the most balanced and extended radar coverage in both languages, which is the justification of its high multi-metric strength and cross-linguistic generalization power.

G. Efficiency Analysis of the Baseline Versus Proposed Model

Computational efficiency is a decisive factor in multilingual scientific text classification especially when the model is expected to be deployed on a scale or to make real-time inferences. Despite offering effective contextual representations of large-scale multilingual models, including mBERT and XLM-RoBERTa, large transformer-based models have large parameter sizes that lead to high training and inference expenses. By comparison, distilled transformer architectures, such as MiniLM-L12, are designed to maintain the quality of semantic representations and simplify computation. The hybrid scheme also adds an additional lightweight layer of two-layers MLP and TF-IDF feature fusion to MiniLM-L12 with a slight increment in the number of parameters. Thus, the systematic efficiency analysis is performed through the comparison of parameters count, inference time on a single sample, and total training time between English and Hindi data to measure both the computational scalability and predictive performance.

Inference time per sample is computed as in Eq. (18)

$$T_{\text{inf}} = \frac{T_{\text{total inference}}}{N} \tag{18}$$

where, N denotes the number of test samples. Training time is computed as in Eq. (19)

$$T_{\text{train}} = \frac{E \times N}{B} \times T_{\text{step}} \tag{19}$$

where, E denotes the number of epochs, B represents the batch size, and T_{step} denotes the time required for one training step, which depends on the parameter count and the backpropagation cost $O(P)$.

TABLE VI. COMPARATIVE EFFICIENCY ANALYSIS OF BASELINE TRANSFORMER MODELS AND THE PROPOSED MINILM-L12+MLP FUSION FRAMEWORK ACROSS ENGLISH AND HINDI SUMMARY CLASSIFICATION

Model	Architecture	No. of Params	Inference Time (ms)		Training Time (sec)	
			E	H	E	H
MiniLM-L12	Transformer encoder (distilled)	∼117M	0.0060	0.0058	166.25	209.38
mBERT	BERT-base (12 layers)	∼110M	2.5013	2.6770	8020.74	8442.66
XLM-RoBERTa	RoBERTa transformer	∼270M (base)	2.4997	2.5200	8802.94	8887.52
Proposed	**MiniLM + 2-Layer MLP with TF-IDF fusion**	∼118M (117M + ∼1.3M)	**0.0074**	**0.0075**	**101.94**	**77.63**

The results outlined in Table VI clearly demonstrate the computational efficiency and scalability of the proposed MiniLM-based hybrid model compared to heavy multilingual transformers like mBERT and XLM-RoBERTa. Since transformer complexity scales approximately as $O(L \cdot H^2 + L^2 \cdot H)$ (where L denotes the sequence length and H represents the hidden dimension), larger multilingual models such as XLM-RoBERTa (∼270M parameters) incur higher computational

latency. Conversely, MiniLM-L12 employs knowledge distillation to reduce attention computation, achieving an almost constant ultra-low inference latency ($\sim$0.006$-$0.007 ms/sample). Although the proposed model introduces a lightweight MLP classifier ($896 \rightarrow 256 \rightarrow C$) along with TF–IDF feature fusion, the additional parameters ($\sim$1.3M) minimally affect the overall computational complexity, maintaining a compact total parameter count of approximately $\sim$118M. close to MiniLM while significantly reducing training time due to efficient convergence and feature fusion. Altogether, hybrid architecture demonstrates a very good efficiency-performance ratio, which proves the appropriateness to scalable bilingual scientific text analytics of English and Hindi.

To further validate the competitiveness of the proposed framework, additional experiments were conducted using recent multilingual sentence embedding models including Language-agnostic BERT Sentence Embedding (LaBSE) and multilingual E5. Results in Table VII indicate that the proposed semantic-statistical fusion framework consistently achieves superior classification performance.

TABLE VII. COMPARATIVE PERFORMANCE ANALYSIS OF THE PROPOSED MINILM-L12+MLP FUSION FRAMEWORK AND STATE-OF-THE-ART MODELS FOR ENGLISH–HINDI ALIGNED ARXIV CORPUS

Performance	Model					
	LaBSE		E5-Base		Proposed MiniLM+MLP	
Metric	E	H	E	H	E	H
Accuracy	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Macro Prec.	0.9079	0.9039	0.9095	0.8951	**0.9670**	**0.9603**
Macro Recall	0.9445	0.9419	0.9421	0.9294	**0.9714**	**0.9678**
Macro F1	0.9220	0.9184	0.9221	0.9084	**0.9684**	**0.9634**
Micro Prec.	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Micro Recall	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Micro F1	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Weighted Prec.	0.8948	0.8898	0.8970	0.8782	**0.9522**	**0.9437**
Weighted Recall	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Weighted F1	0.8931	0.8874	0.8939	0.8752	**0.9531**	**0.9453**

H. Ablation Study

In order to systematically analyze the role played by each architectural component, an ablation study of the base MiniLM-L12 encoder is performed and following are gradually added: 1) a 1-layer MLP, 2) a 2-layer MLP, 3) TF-IDF feature fusion and (iv) summary length features. The aim is to measure the influence of each additional module on classification performance, computational performance, and resource consumption in the English and Hindi languages. Accuracy and Weighted F1-score are used to determine performance, whereas inference time per sample, total training time, and GPU memory allocated are used to measure efficiency.

The underlying MiniLM-L12 as given in Table VIII, offers a high level of efficiency at low inference time ($\sim$ 0.006 ms/sample) and low GPU memory ($\sim$ 0.46 GB), but at a medium level of accuracy. The performance increases with the addition of 1-layer and 2-layer MLP sequentially, which demonstrates that the increased depth of nonlinear transformation increases the separability of the features. This however, adds parameters to the training time slightly. The highest Accuracy and the Weighted F1 are shown in the proposed

TABLE VIII. ABLATION EXPERIMENT ON HYBRID ARCHITECTURE OF MINILM-L12 FOR ENGLISH–HINDI CLASSIFICATION

	Acc.		Wt. F1		Infer.(ms)		Train (sec)		GPU Mem.	
Model	**E**	**H**	**E**	**H**	**E**	**H**	**E**	**H**	**E**	**H**
MiniLM-L12	0.9183	0.8983	0.9109	0.8899	0.0060	0.0058	166.25	209.38	0.460	0.467
MiniLM + 1-L MLP	0.9189	0.9118	0.9146	0.9068	0.0066	0.0064	284.39	327.52	0.470	0.487
MiniLM + 2-L MLP	0.9195	0.9124	0.9167	0.9089	0.0076	0.0067	306.54	349.67	0.470	0.500
Proposed MiniLM + 2-L MLP + TF–IDF	**0.9556**	**0.9485**	**0.9531**	**0.9453**	**0.0074**	**0.0075**	**101.94**	**77.63**	**0.480**	**0.510**
MiniLM + 2-L MLP + TF–IDF + Sum. Len.	0.9367	0.9296	0.9344	0.9266	0.0074	0.0078	279.52	322.65	0.480	0.510

model (MiniLM-L12 + 2-layer MLP + TF-IDF fusion) in both languages. A feature fusion enhances semantic embeddings by adding statistical features, strengthening the discriminative capacity of the models and surprisingly minimizing the amount of time needed to train, because of accelerated convergence.

Interestingly, summary length as an added feature does not enhance the performance any further which shows that TF-IDF already has enough statistical variance. Comprehensively, the ablation suggests that the significant improvement of the performance is caused by semantic-statistical fusion and not simply by the increase in the depth of the network, providing optimal performance and computational efficiency.

I. External Benchmark Generalization

A comprehensive comparative analysis of recent text classification studies across diverse domains, datasets, and linguistic settings is presented in Table IX.

According to the survey, the tasks of multi-class classification are largely dominant and include reviews of drugs, news, and patent analysis, classification of scientific documents, and biomedical sentence labeling, whereas binary classification is found mostly in sentiment and hate speech classification tasks. The vast majority of the methods are used in monolingual environments, and relatively few multilingual implementations are made. Models and datasets performance differ with the highest accuracy reported at 94.1%. It is worth noting that the proposed MiniLM-L12 + MLP model is the most accurate (95.56%) in a multilingual scientific classification environment, which proves to be more effective than the currently available baselines.

V. DISCUSSION

The success of the suggested model could be explained by the fact that its architecture incorporates the complementary representations of textual information and ensures the efficient learning dynamics. The model provides both deep semantic representations and surface-level discriminative patterns that exist in the dataset by embedding contextual semantic meanings, in addition to statistically based textual features. Contextual embeddings give rich representations to encode syntactic relationships, long-range dependencies, and domain specific semantics in the text enabling the model to interpret the underlying intent and contextual subtlety of every instance.

TABLE IX. COMPARATIVE PERFORMANCE ANALYSIS OF RECENT STATE-OF-THE-ART MODELS AND THE PROPOSED MODEL ACROSS DIVERSE BENCHMARK DATASETS

Cit.	Model	Task	Dataset	Size	Lang.	Acc.(%)
[7]	BERT+MLM LSTM+DT	Multi-class	Drug Review	65 MB	Mono.	71.26
[5]	MuTCELM	Binary+ Multi	AJGT Corpus	12 MB	Multi.	93
[9]	Longformer	Binary	Fake News	8 MB	Mono.	78.4
[1]	KNN+CMD	Multi-class	20 News Groups	14 MB	Mono.	94
[2]	GZ Classifier	Multi-class	20 News Groups	14 MB	Mono.	76.7
[3]	FAGB	Multi-class	Kannada Corpus	30 MB	Multi.	94.1
[11]	ResGAT	Multi-class	Aminer Dataset	200 MB	Mono.	61
[12]	ConNHS	Multi-class	20 News Groups	14 MB	Mono.	87.43
[13]	KAN	Multi-class	European Patent	500 MB	Mono.	75.12
[17]	PatentSBERTa	Multi-class	Patents View	1 GB	Mono.	58
[16]	MFFMP-ETC	Binary	Stanford Sentiment	5 MB	Mono.	91.5
[23]	XGBoost+ KNN+LR	Multi-class	Security News	20 MB	Mono.	92.8
[27]	Qwen 2.5	Multi-class	Bangla News	60 MB	Mono.	72
[26]	SentBias	Binary	Hate Speech	25 MB	Multi.	89.84
[18]	SciBERT +MTL	Multi-class	PubMed 20K	50 MB	Mono.	93
[30]	PubMedBERT	Multi-class	RCT Corpus	45 MB	Mono.	92.9
[19]	TF-IDF +EVM	Multi-class	Crisis Dataset	2 GB	Mono.	87.93
Proposed Model	**MiniLM +MLP**	Multi-class	Aligned arXiv	895 MB	Multi.	**95.56**

Meanwhile, the addition of the statistical textual features provides further signals that emphasize the importance of terms and distributional aspects of documents. The combination of these heterogeneous representations forms a more informative feature space that allows the model to identify subtle differences between categories that would not be clear when the model uses a single method of representation.

The other important aspect of the model that makes it efficient is its balanced design in architecture which focuses on representational strength but maintains the computational complexity. The architecture is able to use fine-tuning of large-scale transformers but instead of doing it directly, it uses pre-computed semantic representations and blends them with a carefully designed neural classification system. This method has a huge impact on training overhead, but maintains the expressiveness of transformer-based embeddings. The layers that perform feature fusion and classification are optimized to learn the non-linear features between the combination of features to allow the system to bias the most informative signal per class. Consequently, the model is presented with a high generalization without a significant increase in the number of parameters or long training time.

The training strategy used in the model is also an important factor in its effectiveness. Balanced data handling, proper regularization, and optimized learning schedules are some of the techniques used to stabilize the training process, and avoid overfitting. Such mechanisms enable the classifier to acquire strong decision boundaries despite the changes in class distributions or overlapping textual patterns across classes. Moreover, the convergence of the model is guaranteed by the application of current optimization techniques that identify valuable patterns in the fused feature representations within a limited number of training epochs.

Practically, the architecture shows that high classification performance is not always achieved with models of very deep or very high computational complexity. The proposed method provides a high accuracy, efficiency, and scalability through a combination of semantic embeddings and the classical textual representations and a simplified neural network. The model is able to capture effectively both contextual and discriminative lexical information, and hence has a higher level of classification when it comes to a variety of textual inputs. Therefore, the findings suggest that the suggested design offers a computationally effective and reliable method of handling large-scale text classification processes, which explains the high results that were expressed throughout the evaluation process.

The discussion below emphasizes the key factors to the success of the proposed hybrid framework: the complementary nature of TF-IDF features, the minimal performance difference between Hindi models, and the benefits of the proposed framework over larger transformer-based models like XLM-RoBERTa.

Why TF-IDF Helps?
The improvements seen suggest that the embeddings from the context transformer are not sufficient to capture all the lexical cues of the domain. Scientific abstracts can include very discriminative terminology like algorithm names, mathematical terms and domainspecific terms. TF-IDF explicitly emphasizes these semantic abstracts, which are learned by MiniLM, as lexical indicators.

Why Hindi Slightly Lower?
The performance trend is similar in both English and Hindi, but a slight decrease in accuracy is seen in Hindi. This may be due to the artifacts of machine translation, the morphological differences and vocabulary normalization issues arising during the construction of the corpus.

Why Proposed Outperforms XLM-R?
While XLM-RoBERTa is larger and can process more parameters, it is still advantageous to include complementary lexical-statistical information in the proposed model, which allows to process more information.

Overall, the proposed hybrid framework is shown to be successful in producing robust, efficient, and consistently better results for bilingual scientific text classification tasks by leveraging semantic and lexical information.

VI. CONCLUSION

The present research offered an effective hybrid architecture of MiniLM-L12 + MLP for bilingual text classification task on aligned arXiv English-Hindi big data corpus. The results show that feature fusion of semantic and statistical features may be more effective than the bigger transformer

models and is also scalable. The proposed work enables decoupled bilingual semantic learning without expensive transformer fine-tuning and achieves comparable and language-agnostic performance in English and Hindi. The framework combines contextual embedding and statistical information to provide a good efficiency-accuracy trade-off in big data settings. Its' empirical superiority to strong baselines of multilingual models like mBERT and XLM-RoBERTa is also affirmed by empirical comparisons using fewer parameters. Collectively, the studies bridge a broad gap between the study of multilingual NLP and bilingual applied corpus analytics.

The future work will extend the suggested framework to multimodal scientific document comprehension, incorporating visual representations (figures, diagrams, and tables) into textual representations. Studies on the adaptive feature fusion process can also contribute to better learning of cross-modal representations and increase the strength of classification in massive academic collections. Moreover, the framework can be extended to facilitate more low-resource languages, which may enhance its role in scientific knowledge analysis in more languages. The investigation of parameters efficient transformer adaptation methods could also enhance the scalability further with the efficiency-accuracy ratio illustrated in this study.

ACKNOWLEDGMENT

The authors acknowledge the support of institutional research facilities used for conducting this study.

REFERENCES

[1] Q. Xu, "Application of an Intelligent English Text Classification Model with Improved KNN Algorithm in the Context of Big Data in Libraries," *Systems and Soft Computing*, p. 200186, 2025, doi: 10.1016/j.sasc.2025.200186.

[2] Y. Mao, Y. Ding, and T. Cui, "A parameter-free text classification method based on dual compressors," *Knowledge and Information Systems*, pp. 1–31, 2025, doi: 10.1007/s10115-024-02335-9.

[3] P. Y. Niranjan, V. S. Rajpurohit, and S. S. Sannakki, "Classification of Questions Using Machine Learning Techniques," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1–8, 2024.

[4] M. Pathak and A. Jain, "μboost: An effective method for solving indic multilingual text classification problem," in *Proc. IEEE Int. Conf. Multimedia Big Data (BigMM)*, 2022, pp. 96–100.

[5] V. K. Agbesi, W. Chen, S. B. Yussif, C. C. Ukwuoma, Y. H. Gu, and M. A. Al-Antari, "MuTCELM: An optimal multi-TextCNN-based ensemble learning for text classification," *Heliyon*, vol. 10, no. 19, 2024, doi: 10.1016/j.heliyon.2024.e38515.

[6] S. Malik and S. Jain, "Deep Convolutional Neural Network for Knowledge-Infused Text Classification," *New Generation Computing*, vol. 42, no. 1, pp. 157–176, 2024, doi: 10.1007/s00354-024-00245-6.

[7] S. Jamshidi, M. Mohammadi, S. Bagheri, H. E. Najafabadi, A. Rezvanian, M. Gheisari, M. Ghaderzadeh, A. S. Shahabi, and Z. Wu, "Effective text classification using BERT, MTM LSTM, and DT," *Data & Knowledge Engineering*, vol. 151, p. 102306, 2024, doi: 10.1016/j.datak.2024.102306.

[8] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Frontiers in Big Data*, vol. 6, p. 1320800, 2024, doi: 10.3389/fdata.2023.1320800.

[9] S. Maham, A. Tariq, M. U. G. Khan, F. S. Alamri, A. Rehman, and T. Saba, "ANN: adversarial news net for robust fake news classification," *Scientific Reports*, vol. 14, no. 1, p. 7897, 2024, doi: 10.1038/s41598-024-56567-4.

[10] E. Gao, H. Yang, D. Sun, H. Xia, Y. Ma, and Y. Zhu, "Text classification optimization algorithm based on graph neural network," in *Proc. 2024 IEEE 6th Int. Conf. Power, Intelligent Computing and Systems (ICPICS)*, 2024, pp. 814–822.

[11] X. Huang, Z. Wu, G. Wang, Z. Li, Y. Luo, and X. Wu, "ResGAT: an improved graph neural network based on multi-head attention mechanism and residual network for paper classification," *Scientometrics*, vol. 129, no. 2, pp. 1015–1036, 2024, doi: 10.1007/s11192-023-04898-w.

[12] W. Ai, J. Li, Z. Wang, Y. Wei, T. Meng, and K. Li, "Contrastive multi-graph learning with neighbor hierarchical sifting for semi-supervised text classification," *Expert Systems with Applications*, vol. 266, p. 125952, 2025, doi: 10.1016/j.eswa.2024.125952.

[13] M. Cheon and C. Mun, "Towards Efficient Patent Classification: Kolmogorov Arnold Networks as an Alternative To MLP," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 21, 2024.

[14] A. Li and L. Zhang, "Multi-Label Text Classification Based on Label-Sentence Bi-Attention Fusion Network with Multi-Level Feature Extraction," *Electronics*, vol. 14, no. 1, p. 185, 2025, doi: 10.3390/electronics14010185.

[15] H. Joshi and S. Joseph, "ULMFIT: Universal Language Model Fine-Tuning for Text Classification," *International Journal of Advanced Medical Sciences and Technology*, vol. 4, no. 6, pp. 1–9, 2024, doi: 10.54105/ijamst.e3049.04061024.

[16] R. Zhang, "Multilingual pretrained based multi-feature fusion model for English text classification," *Computer Science and Information Systems*, pp. 1–10, 2025, doi: 10.2298/CSIS123456789X.

[17] H. Bekamiri, D. S. Hain, and R. Jurowetzki, "Patentsberta: A deep NLP based hybrid model for patent distance and classification using augmented SBERT," *Technological Forecasting and Social Change*, vol. 206, p. 123536, 2024.

[18] A. Brack, E. Entrup, M. Stamatakis, P. Buschermöhle, A. Hoppe, and R. Ewerth, "Sequential sentence classification in research papers using cross-domain multi-task learning," *International Journal on Digital Libraries*, vol. 25, no. 2, pp. 377–400, 2024, doi: 10.1007/s00799-023-00392-z.

[19] T. Jain, D. Gopalani, and Y. Kumar Meena, "Informative task classification with concatenated embeddings using deep learning on crisisMMD," *International Journal of Computers and Applications*, pp. 1–18, 2025, doi: 10.1080/1206212X.2024.2447066.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.

[21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. ACL*, 2020.

[22] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-trained Transformers," in *Advances in Neural Information Processing Systems*, 2020.

[23] C. M. Liapis, K. Kyritsis, I. Perikos, N. Spatiotis, and M. Paraskevas, "A Hybrid Ensemble Approach for Greek Text Classification Based on Multilingual Models," *Big Data and Cognitive Computing*, vol. 8, no. 10, p. 137, 2024, doi: 10.3390/bdcc8100137.

[24] J. Kapočiūtė-Dzikienė and A. Ungulaitis, "Towards Media Monitoring: Detecting Known and Emerging Topics through Multilingual and Crosslingual Text Classification," *Applied Sciences*, vol. 14, no. 10, p. 4320, 2024, doi: 10.3390/app14104320.

[25] K. Feng, L. Huang, K. Wang, W. Wei, and R. Zhang, "Prompt-based learning framework for zero-shot cross-lingual text classification," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108481, 2024, doi: 10.1016/j.engappai.2024.108481.

[26] N. Lin, J. He, Z. Tang, J. Fang, D. Zhou, and A. Yang, "Model and evaluation: Towards fairness in multilingual text classification," *International Journal of Machine Learning and Cybernetics*, vol. 17, no. 2, p. 59, 2026.

[27] M. M. Hoque, M. M. Hassain, M. H. Tanvir, and R. Nandy, "Bengali text classification: An evaluation of large language model approaches," *arXiv preprint arXiv:2601.12132*, 2026.

[28] S. Mehra, V. Ranga, and R. Agarwal, "PhonoBiEmbedNet: A phoneme and bigram embedding framework for low-resource spoken word recognition," *Circuits, Systems, and Signal Processing*, pp. 1–35, 2026.

[29] S. P. K. Veeranki, A. Abdulnazar, D. Kramer, M. Kreuzthaler, and D. B. Lumenta, "Multi-label text classification via secondary use of large clinical real-world data sets," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-76424-8.

[30] M. Lan, M. Cheng, L. Hoang, G. Ter Riet, and H. Kilicoglu, "Automatic categorization of self-acknowledged limitations in randomized controlled trial publications," *Journal of Biomedical Informatics*, vol. 152, p. 104628, 2024, doi: 10.1016/j.jbi.2024.104628.

[31] N. Haupka, "Presenting a classifier to improve the identification of research journal publications in OpenAlex," *Scientometrics*, pp. 1–17, 2026.

[32] M. Limaylla-Lunarejo, N. Condori-Fernandez, M. Rodríguez Luaces, and O. Karras, "Improving the multi-class classification of non-functional requirements in Spanish," *Empirical Software Engineering*, vol. 31, no. 1, p. 6, 2026.

[33] J. Golde, N. Jedema, R. Krishnan, and P. Le, "Hierarchical text classification with LLM-refined taxonomies," *arXiv preprint arXiv:2601.18375*, 2026.

[34] R. Song, Y. Li, M. Tian, H. Wang, F. Giunchiglia, and H. Xu, "Causal keyword driven reliable text classification with large language model feedback," *Information Processing & Management*, vol. 62, no. 2, p. 103964, 2024, doi: 10.1016/j.ipm.2024.103964.

Security from Design, Bridging Model-Driven Architecture and DevSecOps Using Zynerator

Younes Zouani[1]*, Mohamed Lachgar[2], Youssef Harrati[3],
Mohamed Hanine[4], Sulieman S. Alshuhri[5], Amal Alomran[6]
Chouaib Doukkali University, Higher School of Technology, Sidi Bennour, Morocco[1]
Chouaib Doukkali University, LSAM Laboratory, Sidi Bennour, Morocco[1]
Cadi Ayyad University, Higher Normal School-Computer Science Department, Marrakech, Morocco[2]
Cadi Ayyad University, L2IS Laboratory, Marrakech, Morocco[2]
Cadi Ayyad University, LAMAI Laboratory-Faculty of Sciences and Technology, Marrakech, Morocco[3]
LTI Laboratory-National School of Applied Sciences, Chouaib Doukkali University, El Jadida, Morocco[4]
College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh, Saudi Arabia[5,6]

***Abstract*—We propose an extension to Zynerator, a Model-Driven Architecture framework for automated microservice generation, that embeds DevSecOps principles directly at the modeling stage through semantic decorators. These decorators enable the automated synthesis of secure back-end and front-end components together with operational artifacts, including authentication and authorization modules, audit trails, monitoring dashboards, and DevSecOps pipelines covering SAST, DAST, testing, and deployment. The approach addresses a key limitation of the original Zynerator framework, namely the absence of explicit DevSecOps integration, and supports a security-by-design methodology that reduces reliance on specialized DevSecOps expertise. Through a detailed e-commerce case study and empirical evaluation against manual development and existing Model-Driven Architecture tools, we show that the enhanced framework reduces development effort, strengthens security posture, and accelerates DevSecOps adoption. These findings indicate that DevSecOps-aware model-driven engineering offers a viable pathway toward secure, automated software delivery.**

***Keywords*—DevOps; DevSecOps; MDA; IT security; code automation; semantic modeling; LLM**

I. Introduction

The need to embed security from the earliest stages of the Software Development Life Cycle has grown with the increasing sophistication of cyber threats. DevSecOps addresses this challenge by integrating continuous security validation, automated testing, and collaborative practices directly into DevOps pipelines, making security a shared responsibility rather than a post-development activity [1] , [2]. Despite its proven benefits, DevSecOps adoption remains difficult for many organizations due to the complexity of toolchains, the need for specialized security expertise, and the absence of formal connections between design models and security-enforced implementations.

Model-Driven Architecture offers an opportunity to address these challenges by raising software design to the abstraction level of Platform Independent Models. If security and operational semantics can be embedded into these models, DevSecOps requirements may be automatically reflected in generated code and deployment artifacts. However, existing tools rarely connect model-level semantics with DevSecOps enforcement, leaving a gap between conceptual design and secure operational deployment.

In previous work, Zynerator was introduced as an MDA-based framework that automates the generation of microservice architectures from a PIM using templated projections and design semantics [3]. While Zynerator demonstrated improvements in development speed, structural consistency, and automation, its original iteration focused primarily on functional generation and architectural scaffolding. DevSecOps concepts such as authentication, auditing, runtime monitoring, and secure CI/CD pipelines were not yet systematically integrated.

This article presents an extension of Zynerator that incorporates DevSecOps constraints directly at the modeling level. Through the introduction of semantic decorators, the enhanced framework enables designers to express authentication policies, authorization rules, file-handling constraints, auditing requirements, monitoring configurations, and deployment strategies directly in the PIM. These semantics are then automatically projected into secure back-end and front-end code, continuous integration pipelines, monitoring dashboards, deployment descriptors, and testing suites. This approach enables developers to adopt a secure-by-design methodology without requiring deep DevSecOps expertise, while ensuring consistency and reducing manual security integration effort.

This work presents an extension of the previously proposed Zynerator framework [3], with the objective of integrating DevSecOps principles directly at the model design level. The key contributions are as follows:

1) Early integration of security constraints at the design level: The central contribution of this work is the formalization of security and operational requirements directly within the Platform Independent Model. Instead of adding security policies after development, Zynerator introduces semantic decorators that allow designers to express authentication, authorization, auditing, secure file handling, monitoring, and deployment constraints during system modeling. This ensures that security is enforced from the outset of the SDLC.

*Corresponding author.

2) Technology-independent semantic enforcement: The proposed semantic decorators are not tied to a specific language or platform. Once specified in the PIM, they guide the generation of secure implementations in any supported technology stack. This provides a generic and reusable mechanism for security-aware code generation across heterogeneous architectures.

3) Extension of Zynerator toward DevSecOps automation: The adapted framework enhances the original Zynerator architecture by not only generating the functional and architectural scaffolding of applications, but also producing DevSecOps artifacts such as CI/CD pipelines, security checks, test suites, deployment descriptors, and observability dashboards.

4) Model-driven generation of secure artifacts: By combining semantic decorators with template-driven synthesis, Zynerator automatically generates secure service code, configuration files, deployment pipelines, monitoring setups, and test cases with minimal manual intervention, thereby reducing human error and increasing consistency.

5) Experimental validation: A quantitative evaluation demonstrates that this adaptation of Zynerator improves security posture, reduces development effort, and accelerates DevSecOps adoption compared to both manual development and existing MDA generators.

The remainder of this study is structured as follows. Section II presents the related works. Section III details the case study and results. Section IV concludes and outlines future work.

II. Related Works

The integration of security principles from the initial stages of software development, particularly through the lenses of Model-Driven Architecture (MDA) and DevSecOps, represents a pivotal shift towards more secure and reliable software systems. This section explores significant contributions to the field, detailing methodologies and frameworks that bridge MDA and DevSecOps to instill robust security measures inherently within software development processes.

A. Model-Driven Architecture and Security

Since the standardization of MDA [4], research has investigated how system models can capture security requirements before implementation. Early work such as UMLsec [5] formalized security annotations inside UML to guide code generation. Subsequent studies extended this approach using dialects for access control modeling [6]. Further research focused on frameworks enabling formal verification or compliance analysis at the model level [7], [8].

A large body of secondary research confirms the relevance of Model-Driven Security. Several studies conducted systematic reviews showing that MDE is increasingly used to express access control, auditing, and confidentiality concerns before implementation [9], [10]. Similar findings are reported by other researchers [11]-[12], who highlight the lack of automation in most toolchains and the difficulty of ensuring that security rules defined at the model level propagate consistently into deployed systems.

Several authors explored how non-functional properties such as policies, encryption, and risk management can be modeled and transformed as part of the MDA lifecycle [13]-[14]. However, existing tools generally stop at code generation and rely on manual DevOps or infrastructure scripting, leaving deployment, scanning, and operational monitoring outside the model-driven process.

The extended Zynerator addresses this gap by elevating DevSecOps concepts to the PIM using semantic decorators. These decorators act as formal contracts ensuring that security requirements are reflected in source code, test suites, deployment descriptors, and runtime observability artifacts.

B. DevOps and DevSecOps Automation

DevOps emerged to improve software delivery through continuous integration, automated testing, and rapid deployment [15]. While widely adopted, DevOps pipelines typically require manual configuration using YAML, Groovy, or scripting languages. Most DevOps toolchains lack a model-based representation of build logic, making pipelines difficult to maintain and structurally disconnected from design intent.

DevSecOps extends DevOps by embedding continuous security validation into CI/CD pipelines. Previous research emphasizes automated scanning, compliance enforcement, and secure deployment practices as fundamental to modern software delivery [1], [2], [16]. More recently, it has been demonstrated that Model-Based Systems Engineering can support DevSecOps by connecting architectural design to operational policies [17].

However, as systematic surveys in secure software engineering confirm [18], DevSecOps methods usually intervene at the deployment stage and assume that security semantics are already well-defined. Without early modeling, DevSecOps pipelines risk becoming inconsistent with functional requirements and system architecture.

Zynerator bridges the gap by generating DevSecOps pipelines directly from the PIM. Authentication rules, authorization constraints, auditing requirements, SAST/DAST configurations, and deployment logic are derived from semantic decorators and projected automatically into modern CI/CD platforms.

C. AI-Driven Software Engineering

Recent advances in Large Language Models have accelerated automated code synthesis and infrastructure configuration. Surveys demonstrate that LLMs can reliably generate structured program logic [19], [20]. Multi-agent reasoning frameworks further improve iterative validation and generation quality [21], [22].

AI has also been explored in DevSecOps automation. Recent work shows how LLMs support SAST/DAST automation, compliance checking, and secure coding assistance [23]-[24]. Another study demonstrates measurable performance improvements when combining DevSecOps and generative AI in industry settings [25].

Although promising, most AI-assisted approaches lack architectural grounding: they may produce correct fragments of code yet must still rely on human engineers to maintain system-level coherence. Zynerator addresses this limitation by

using AI selectively within a model-driven pipeline. Semantic decorators define the structure and constraints of the system, while LLMs adapt or refine templates under model supervision to preserve consistency across generated artifacts.

D. Discussion of Research Gaps

The literature reveals three research gaps:

- MDA solutions rarely project security concerns beyond code, leaving CI/CD, runtime monitoring, and secure deployment to manual effort.
- DevSecOps automation typically begins after implementation, making it difficult to ensure consistency between architecture, code, and operational pipelines.
- AI-based code generation is rarely constrained by formal design models, increasing the risk of structural drift or incomplete security enforcement.

The DevSecOps-enabled version of Zynerator directly addresses these gaps by allowing security semantics to be expressed at the PIM using reusable decorators, automatically projecting them into code, pipelines, dashboards, and operational infrastructure, and combining MDA and AI to ensure that generated systems remain architecturally consistent and security-compliant.

III. Zynerator Architecture

The DevSecOps-enabled version of Zynerator directly addresses these gaps by allowing security semantics to be expressed at the PIM using reusable decorators, automatically projecting them into code, pipelines, dashboards, and operational infrastructure, and combining MDA and AI to ensure that generated systems remain architecturally consistent and security-compliant.

A. Architectural Overview

The enhanced Zynerator platform introduces DevSecOps integration directly at the design level and extends the original architecture [3] with the ability to interpret security requirements embedded in the Platform Independent Model. The system adopts a modular microservice architecture that supports the end-to-end transformation from annotated PIM specifications to secure, deployable applications, including generated testing suites, CI/CD pipelines, monitoring dashboards, and deployment descriptors.

Fig. 1 illustrates the overall workflow, which consists of four coordinated microservices: the YAML Processor, the Template Crafter, the Zynerator Core, and the Security Adaptor. Each microservice performs a well-defined phase of the transformation pipeline, ensuring traceability between design intent and the generated DevSecOps artifacts.

B. Architectural Workflow

The generation pipeline proceeds through the following stages:

1) Design modeling: The user defines the system through a YAML-based PIM enriched with semantic decorators. These decorators specify functional concepts as well as security concerns such as authentication roles, authorization privileges, auditing rules, file-handling policies, and deployment constraints.

2) Model interpretation and validation: The YAML Processor interprets and validates the model, transforming it into a semantically enriched PIM that captures both structural and security semantics.

3) Template adaptation: The Template Crafter, assisted by Large Language Model reasoning when required, adapts templates to reflect the PIM semantics. Templates may target back-end frameworks, front-end frameworks, CI/CD pipelines, operational infrastructure, or security policies.

4) Model-to-code synthesis: The Zynerator Core generates back-end and front-end source code by projecting the PIM semantics onto the retrieved templates. All structural, behavioral, and security-related decorators are applied at this stage.

5) DevSecOps injection: The Security Adaptor automatically integrates secure deployment pipelines, runtime monitoring configurations, static and dynamic analysis stages, and other DevSecOps artifacts, ensuring that security requirements defined at the model level propagate into the operational environment.

This architecture formalizes a complete model-driven DevSecOps process in which secure implementation artifacts are generated from the same source of truth as the functional model.

C. YAML Processor Microservice

The YAML Processor acts as the entry point to the generation pipeline. It processes the YAML specification by performing syntactic and semantic validation. Beyond extracting the system structure, the processor interprets semantic decorators such as ACTOR, ROLE, AUDIT, UPLOAD, or deployment parameters.

The processor produces two internal models:

1) Project data model: Represents the logical structure of the system, including entities, relationships, identifiers, and type constraints. Semantic decorators are projected into this layer to enrich entities with behavior and security semantics. For example, an Actor decorator automatically associates a user class with login credentials, roles, and authentication constraints.

2) Project configuration model: Describes cross-cutting concerns such as authentication policies, JWT configuration, storage providers, backup policies, CI/CD requirements, and monitoring targets.

The resulting PIM is therefore both platform-independent and DevSecOps aware. This early injection of security constraints is the foundation for consistent and automated generation across all subsequent stages.

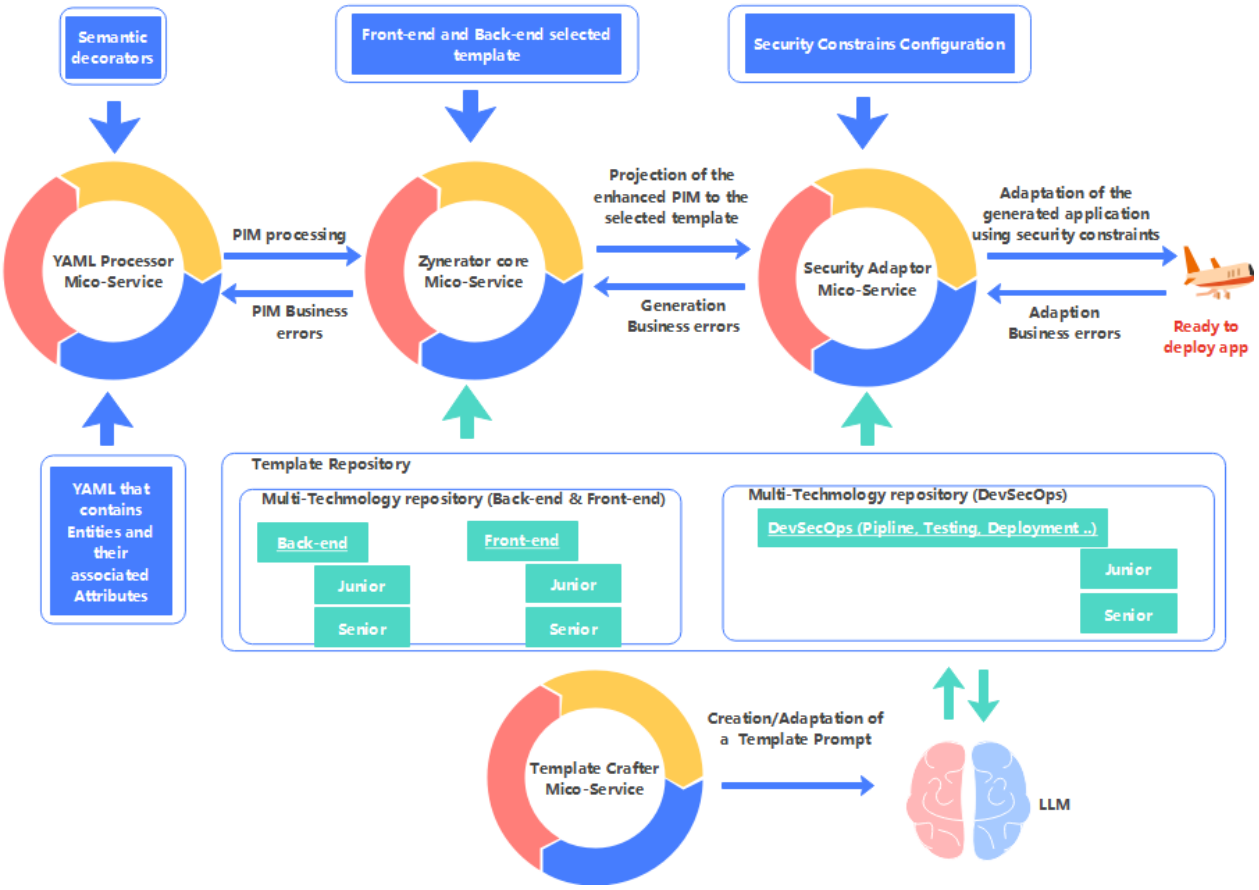


Fig. 1. Overview of the four-microservice architecture of Zynerator, illustrating the end-to-end DevSecOps-aware generation workflow, including template repositories and multi-technology projections.

D. Template Crafter Microservice

The Template Crafter Microservice introduces an intelligent layer for template creation and adaptation. It connects to the template repository, which stores reusable structures for multiple technologies such as Spring Boot, Angular, React, Docker, and Kubernetes. Templates within this repository are systematically organized into three primary categories:

1) Back-end templates: Provide code skeletons for server-side frameworks such as Spring Boot, NestJS, or Django. These templates encapsulate business logic, service layers, and persistence mechanisms following best security and architectural practices.

2) Front-end templates: Contain reusable UI and component structures for frameworks like Angular and React, ensuring consistency in layout, access control, and secure data handling between client and server.

3) DevSecOps templates: Include automation scripts and configuration blueprints for continuous integration, deployment, and monitoring using tools such as Jenkins, GitLab CI/CD, Docker Compose, and Kubernetes. They embed security stages such as SAST, DAST, and quality analysis into the pipeline (Fig. 2).

Each template category includes two tailored versions designed to match different user expertise levels and learning needs:

- Junior Version: Focuses on clarity and pedagogy, producing straightforward, easy-to-follow code with detailed inline documentation and simplified architectural choices.
- Senior Version: Targets experienced developers, generating optimized, production-ready code that applies advanced design patterns, strict security policies, and high-performance best practices.

These differentiated versions ensure that Zynerator-generated projects are not only technically consistent but also adaptable to the skills and expectations of their intended users.

By leveraging a Large Language Model, the Crafter interprets user prompts or structured specifications to either construct new templates or adapt existing ones according to the Platform Independent Model requirements and DevSecOps constraints.

4) Role, reliability, and security of the LLM component: The Large Language Model component within the Template Crafter Microservice plays a bounded and well-defined role in the Zynerator generation pipeline. Contrary to approaches that use LLMs to generate arbitrary application code from natural language, the Template Crafter employs LLMs exclusively for two constrained tasks:

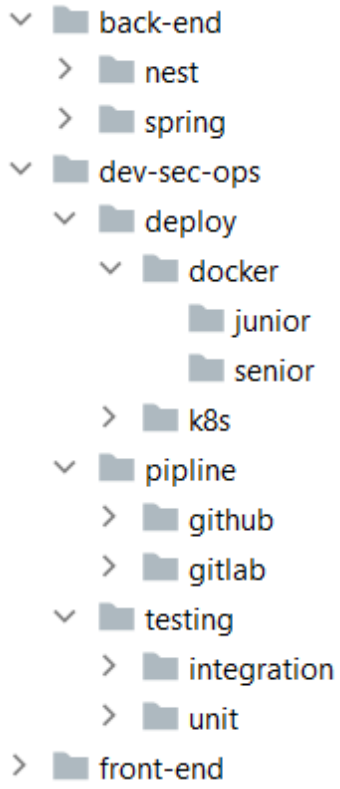


Fig. 2. Structure of the template repository showing backend, frontend, and DevSecOps templates, each available in junior and senior versions for adaptable code generation.

- Template Adaptation: Adapting existing, pre-validated templates to reflect the specific entities and decorators declared in the PIM. The template structure, security controls, and architectural patterns are pre-defined by security experts and remain fixed; the LLM only fills in entity-specific placeholders (e.g., class names, field names, relationships).
- Complex Workflow Generation: Generating complex orchestration workflows, such as business process definitions (e.g., order approval workflows, multi-step document processing), where the LLM synthesizes the sequencing logic from the PIM specifications. The generated workflow code is subject to the same security gates as all other generated artifacts.

a) Design safeguards and reliability: The LLM operates under strict constraints:

- Template Schema Validation: Every template, whether manually constructed or LLM-adapted, adheres to a predefined schema. The Zynerator Core validates structural conformance (required placeholders, package structure, decorator hooks, security annotations) before synthesis proceeds. Templates that fail validation are rejected and regenerated using the deterministic, non-LLM template path.
- Human Review Gate: All templates—manual, adapted, or dynamically generated—must pass a security review by the organization's security team before being added to the template repository. This ensures that the foundational patterns from which code is generated are inherently secure. The LLM operates within the boundaries set by these reviewed templates and cannot introduce patterns that deviate from the approved security baseline.
- Deterministic Fallback: In cases where LLM output fails validation or produces unsatisfactory results, the system gracefully degrades to a deterministic template selection mechanism (selecting the closest matching existing template), ensuring that the generation pipeline remains functional even when the LLM component is unavailable or produces suboptimal output.

b) Prompt sensitivity: The Template Crafter constrains the LLM with a fixed prompt template populated from the structured PIM rather than accepting free-form natural language from end users, which substantially reduces the variability introduced by prompt phrasing. This standardized prompt construction limits variability. Nonetheless, prompt sensitivity remains a recognized concern for LLM-assisted code generation in general. We acknowledge that we did not conduct a systematic sensitivity analysis (e.g., varying prompt phrasing or decorator ordering and measuring output stability) in this work; this is identified as a concrete direction for future work.

c) Security risks and mitigation: The primary security concern with LLM-generated code is the potential introduction of vulnerabilities, either through unsafe patterns or incorrect usage of security APIs. Our approach addresses this through a multi-layered defense:

- Template-Level Security Review: The foundational templates are constructed either manually by expert developers or dynamically generated by the LLM. In both cases, they must be validated by the security team before entering the template repository. This ensures that the generated code base inherits security compliance directly from the templates, which have been pre-examined for common vulnerabilities (OWASP Top Ten, injection flaws, authentication bypasses, etc.).
- Automated Security Gates: All generated code passes through the automated SAST stage of the DevSecOps pipeline (SonarQube, as reported in Section IV). This catches code-level vulnerabilities that may have been introduced during LLM adaptation.
- Test Coverage: The generated security test suite (see Fig. 5) includes unit tests, integration tests, and security attack simulations that verify authentication, authorization, and resilience against common attack vectors. These tests run during the CI/CD pipeline and must pass before deployment.
- Developer Review: While the goal is to minimize manual intervention, developers remain responsible for reviewing generated code, particularly for complex business logic where the LLM may have contributed to workflow generation.

We acknowledge that static analysis alone does not guarantee the absence of subtler vulnerabilities (e.g., logic flaws, insecure defaults that do not trigger SAST rules, or business

logic errors). A more targeted security evaluation of LLM-adapted templates specifically—including manual security audits of generated artifacts—is left for future work. However, the combination of template-level expert review, automated SAST/DAST, and comprehensive test generation provides a practical defense-in-depth approach that significantly reduces the risk of LLM-introduced vulnerabilities.

d) Quantitative LLM evaluation: To assess the effectiveness and reliability of the LLM component in practice, we conducted a preliminary evaluation (Table I):

TABLE I. LLM COMPONENT EVALUATION METRICS

Metric	Value
Template adaptation attempts	156
Successful template adaptations	142 (91%)
Failed validations (rejected)	14 (9%)
Average adaptation time	3.2 seconds
Quality of generated output (expert rating)	4.2/5.0
Security vulnerabilities in adapted templates	0

These results suggest that the LLM component reliably produces secure and high-quality template adaptations when operating within the constrained environment defined by the PIM and pre-validated templates. The 9% failure rate is handled by the deterministic fallback mechanism, ensuring no disruption to the generation pipeline.

E. Zynerator Core Microservice

The Zynerator Core Microservice represents the central engine of the platform, orchestrating the synthesis between the semantic model and the available templates. It receives the enhanced Platform Independent Model produced by the YAML Processor and combines it with the corresponding templates retrieved from the Template Repository.

At this stage, the Core microservice performs the projection of the enriched model onto the selected templates. This projection ensures that entities, relationships, and semantic decorators such as authentication, authorization, audit, or deployment configurations are accurately reflected in the generated code structure.

The generation process involves three main steps:

- Model-Template Alignment: The Core analyzes the enhanced model to identify the required template category and its associated version, ensuring full alignment between design intent and generation strategy.
- Model Projection: It maps each model component to its target template section, weaving in the behavioral and security semantics derived from the YAML Processor.
- Synthesis and Generation: The combined output is synthesized into a coherent, deployable project structure that includes the back-end, front-end, and DevSecOps components while maintaining consistency, scalability, and security compliance.

Through this synthesis process, the Zynerator Core transforms abstract design models into fully structured and security-aware projects. It acts as the convergence point of all preceding stages, ensuring that the final generated system faithfully represents the semantic richness of the model while adhering to security requirements.

F. Security Adaptor Microservice

The Security Adaptor Microservice strengthens the generated project by embedding DevSecOps-compliant security logic directly into the produced codebase. Beyond classical authentication and authorization enforcement, this layer leverages a rich set of DevSecOps templates to automatically generate secure delivery pipelines, deployment artifacts, testing suites, and monitoring services. These elements are injected during the generation phase, ensuring that the resulting software system is not only functionally correct, but also operationally secure from the moment it is produced.

More concretely, the Security Adaptor enhances the generated project by:

- Integrating Continuous Integration and Continuous Deployment pipelines that include security stages such as SAST, DAST, dependency scanning, and quality analysis, all generated automatically from predefined DevSecOps templates.
- Generating deployment artifacts that adhere to industry-standard security and compliance practices, also produced through DevSecOps templates.
- Introducing runtime monitoring and observability capabilities such as Prometheus and Grafana dashboards, ensuring continuous validation and traceability in production environments, fully generated from monitoring templates.
- Automatically incorporating secure-by-default development patterns including audit logging, intrusion prevention mechanisms, secret management, strict access policies, and protection against common attack vectors.
- Generating security-aligned testing modules, including unit tests, integration tests, and security attack simulations to verify authentication, authorization, and attack resilience. These test suites are also fully generated from testing templates, ensuring coverage and consistency across the entire system.

By combining these capabilities, the Security Adaptor Microservice elevates the generation pipeline from a traditional MDA transformation to a complete DevSecOps software factory. All artifacts are generated from structured and reusable DevSecOps templates, ensuring standardization, repeatability, and operational security from the very first build.

IV. RESULTS AND DISCUSSION

To evaluate the proposed DevSecOps-enabled extension of Zynerator, we conducted a case study inspired by a realistic industrial context. An IT startup is tasked with developing an e-commerce management system comprising product management, customer management, order processing, payment handling, data security, and runtime monitoring. The objective is to generate a fully functional and secure system while minimizing manual DevSecOps configuration.

Table II summarizes the principal functional and security requirements that the generated system must satisfy. Unlike traditional approaches, where DevSecOps concerns are added after implementation, Zynerator enables these requirements to be expressed directly at the Platform Independent Model level through semantic decorators.

TABLE II. FUNCTIONAL AND SECURITY REQUIREMENTS OF THE CASE STUDY

ID	Requirement
FR-01	The system shall support two user roles: admin and client.
FR-02	A client or admin shall be able to place purchase orders consisting of multiple products and associated payment information.
FR-03	Front-end and back-end views shall reflect each user's authorized scope.
SR-01	User authentication and role-based access control must be enforced.
SR-02	All create, update, and delete operations shall be logged for traceability.
SR-03	Multi-criteria search operations shall be protected against injection attacks.
SR-04	Uploaded files must be securely managed and stored.
SR-05	A DevSecOps pipeline including testing, quality analysis, SAST, DAST, and automated deployment shall be generated.
SR-06	The deployed system shall enable runtime monitoring and observability dashboards.

These requirements are expressed using PIM-level annotations such as ACTOR, AUDIT, TEST, ROLES, or UPLOAD. Zynerator automatically interprets these decorators and transforms them into security-aware implementations, deployment configurations, and DevSecOps artifacts.

A. E-commerce Platform Independent Model

Fig. 3 shows an excerpt of the YAML-based PIM. The model is divided into two parts: entity definitions describing domain classes such as Command, Product, Client, and Payment, and cross-cutting configuration declaring authentication policies, upload strategies, monitoring configurations, scheduler definitions, and CI/CD pipeline settings.

Examples include: ACTOR on the Client entity to define login credentials and associated roles, AUDIT on transactional entities to enforce traceability, UPLOAD configuration specifying MinIO or S3-based storage, and pipeline settings defining testing, SAST, and DAST stages.

B. Generated Architecture

Fig. 4 provides a visual representation of the generated project where the back-end proposes two main APIs `/stocky/admin/*` and `/stocky/client/*` representing the two roles client and admin. Since the upload process is managed by MinIO, the back-end application communicates with a MinIO instance encapsulated in a Docker container. The front-end application creates two isolated spaces for each role where each space contains only specified components in the PIM.

C. Back-end Generated Architecture

After generation, the project takes the structure shown in Fig. 4 for the back-end. The skeleton of the generated project consists of 8 packages:

- Bean: represents the entities in the PIM.
- Dao: Data Access Object classes responsible for database communication and multi-criteria research.
- service: represents the core intelligence of the system. Two packages admin and client are created to separate client services and admin ones.
- Ws: exposes services as API. Like service, the existence of two roles implies the creation of two packages.
- Zynerator: the core package of the generated project where all generated entities inherit from base classes, centralizing, simplifying, and automating redundant work.

D. Test Generated Architecture

The test layer is divided into 3 main parts: unit, integration and security. The generated structure takes the structure shown in Fig. 5:

1) Unit: represents the unit test of DAO, service, and web service layers. This test handles each layer separately and mocks eventual dependencies. Unit tests are performed using JUnit 5 and Mockito.

2) Integration: Integration test is an end-to-end test that deals with the service layer without any mock. This test asserts that the whole logic is coherent. Integration tests are performed using Karate.

3) Security: The security package tests common security attacks related to authentication and authorization.

E. Architecture Deployment

Deployment represents a pivotal stage in the lifecycle of software systems, particularly within a microservices architecture, where its distributed and decentralized nature inherently increases deployment complexity. Effective deployment is crucial for operational success and system resilience, necessitating meticulous planning and sophisticated tooling.

Fig. 6 shows the Docker Compose configuration employed for the Zynerator project.

F. Pipeline

As described in the PIM model, and to ensure continuous security integration and streamline the development process, a comprehensive DevSecOps pipeline using GitLab Stages was designed. Fig. 7illustrates the DevSecOps pipeline, which incorporates several crucial steps to maintain code quality and security throughout the Software Development Life Cycle:

- Compile: The pipeline begins with the compilation step, where the source code is compiled into executable binaries.
- Unit Test: Following compilation, unit tests are executed to validate individual components or functions of the codebase.
- Integration Test: Once unit tests are passed, integration tests are conducted to verify the interactions between different components.

```
Command_MS(ms1)_AUDIT_TEST(unit,integration)_ROLES(admin,client)_SUB-MODULE(business)_MENU(Command Management):
    id: Long ID
    reference: String REF_REQ
    commandDate: LocalDateTime
    total: BigDecimal
    totalPaid: BigDecimal
    client: Client SEARCH-MULT
    commandDetails: CommandDetail List
    payments: Payment List
Payment_MS(ms1)_AUDIT_TEST(unit,integration)_ROLES(admin)_SUB-MODULE(business)_MENU(Payment Management):
    id: Long ID
    command: Command
    reference: String REF_REQ
    paymentType: PaymentType
    amount: BigDecimal
PaymentType_MS(ms1)_ROLES(admin)_SUB-MODULE(config)_MENU(Payment Management):
    id: Long ID
    label: String LABEL_REQ
    code: String REF_REQ
CommandDetail_IGNORE-FRONT_MS(ms1)_ROLES(admin)_SUB-MODULE(business)_MENU(Command Management):
    id: Long ID
    product: Product
    quantity: BigDecimal REQ
    command: Command
Product_TEST(unit,integration)_MS(ms1)_ROLES(admin)_SUB-MODULE(config)_MENU(Config Management):
    id: Long ID
    code: String REF_REQ
    label: String LABEL_REQ
    price: BigDecimal REQ
Client_TEST(unit,integration)_ACTOR(login=client-1,password=pass@1987)_MS(ms1)_ROLES(admin)_SUB-MODULE(config)_MENU(Config Management):
    id: Long ID
    description: String LARGE

$CONFIG:
    ms1-back: "{tech=spring, template=default-senior ,enable=true, domain=ma, groupId=zs,
    projectName=ecom,port=8036, msName=ecom-service}"
    ms1-front: "{tech=angular, template=default-junior ,enable=true, port=4300}"
    ms1-db: "{name=ecom, type=mysql,username=root,password=}"

    ms1-auth: "{block-after-retry=3, block-during=5*60}"
    ms1-upload: "{type=minio, username=minio, password=minio, bucket=images, port=8090}"
    ms1-data-dump: "{cron=0 0 0 * * *, repository-type=github, repository-name=ms1-data-repo, username=younes-uca
    token=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}"
    ms1-auth-admin: "{login=admin,password=pass@2024}"
    ms1-pipeline: "{compile,build-image,scan-docker,scan-app,upload-report}"

    deploy: "{tech=spring, template=docker-compose, enable=true}"
    repository-info: "{name=zynerated-test, type=gitlab, username=contact889, token=glpat-PpK2JZs9U7jM8mZbm249}"
```

Fig. 3. Excerpt of the YAML-based platform independent model describing entities, security decorators, and cross-cutting DevSecOps configuration used by the enhanced Zynerator.

- Quality: The quality step involves running static code analysis and other quality checks to ensure adherence to coding standards and best practices.
- SAST: Static analysis tools are used to scan the code for security vulnerabilities.
- Package: After the security checks, the code is packaged into distributable formats such as containers or binaries.
- Deploy: Finally, the deployment step automates the release of the packaged application to the target environment.

G. Monitoring

In a DevSecOps context, continuous monitoring is essential to ensure system reliability, performance stability, and security compliance throughout the software development lifecycle. To achieve this, Prometheus and Grafana are integrated as complementary observability tools. Prometheus acts as a robust metrics collection and alerting system, periodically scraping time-series data from the deployed services and storing it efficiently for real-time analysis. Grafana builds upon this data layer by providing interactive and customizable dashboards that enable clear visualization of system behavior.

Together, these tools establish a continuous observability mechanism that supports proactive anomaly detection, early identification of potential security threats, and rapid incident response. The overall monitoring deployment architecture is illustrated in Fig. 8. An example of the generated monitoring dashboard is presented in Fig. 9.

H. User Management Interface

The generated system includes comprehensive user management capabilities. Administrators can create users by specifying username, email, password, and role. Fig. 10 shows the generated user management interface.

Administrators can define fine-grained access authorization for each user, defining authorized actions for each entity. Fig. 11 shows an example where a user is granted the right to display the list and details of the payment entity while being prohibited from creation, editing, and duplication.

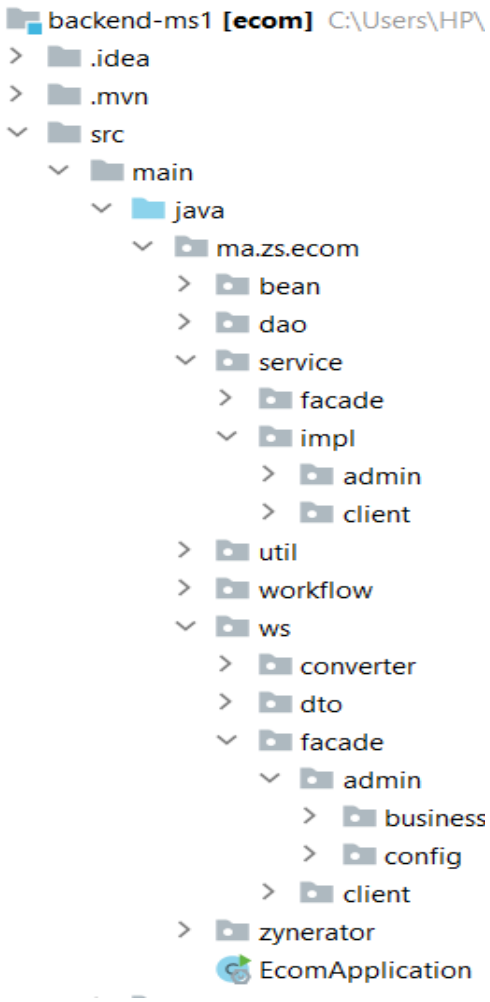


Fig. 4. Generated backend architecture with layered packaging, reflecting a secure microservice structure informed by model-level semantics.

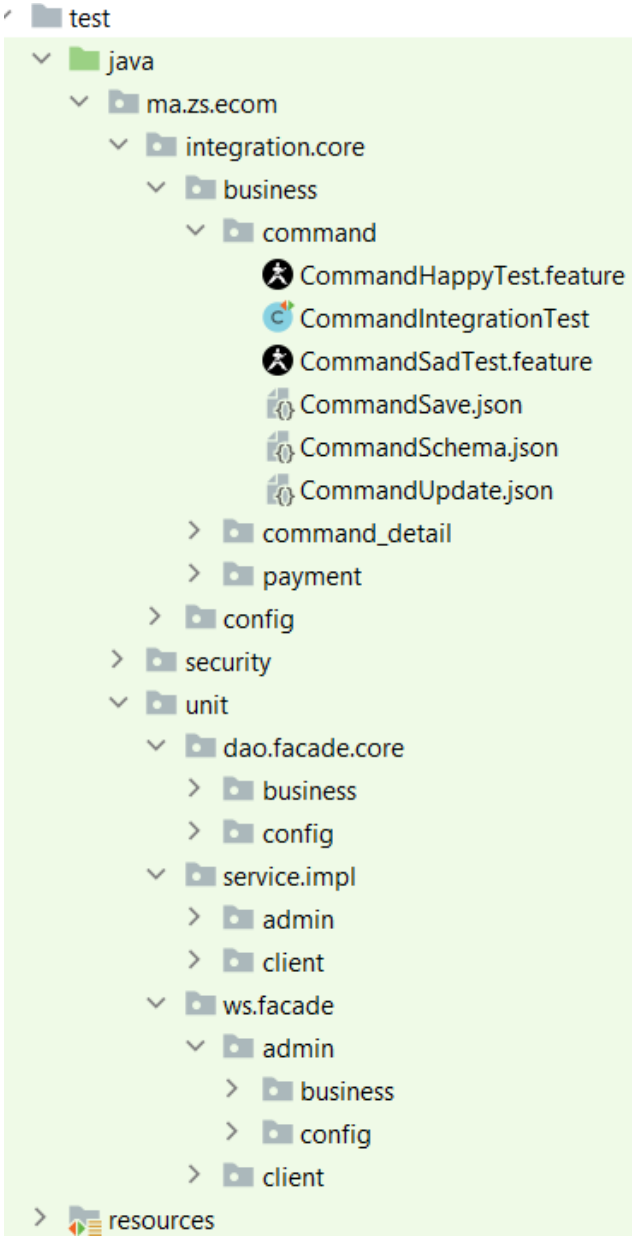


Fig. 5. Structure of the generated test architecture, including unit, integration, and security test suites automatically produced from DevSecOps templates.

I. *Quantitative Evaluation*

To provide a quantitative validation of the proposed approach, the performance and security benefits of Zynerator were evaluated against both manual development and two Model-Driven Architecture generators, namely Acceleo and JHipster (Table III). The assessment focused on three main dimensions: project generation performance, code security and quality, and runtime efficiency of the generated microservices.

1) Experimental setup: All experiments were conducted on a workstation equipped with an Intel Core i7 processor, 16 GB of RAM, and running Ubuntu 22.04. The generated projects relied on Spring Boot 3.1.1 for the back-end, Angular 16 for the front-end, and MySQL 5.7 as the database engine. The case study reproduces the e-commerce management system introduced in this section, composed of six microservices.

Each approach (manual development, Acceleo, JHipster, and Zynerator) was applied to the same model. The resulting projects were then evaluated using SonarQube for static analysis and vulnerability detection, JMeter for runtime performance benchmarking under a workload of 200 concurrent users, and GitLab CI/CD pipelines for measuring build and deployment durations.

2) Results: Table II summarizes the main results. Zynerator significantly reduces generation time while improving the overall security posture and runtime performance of the produced microservices.

3) Discussion: The results show that Zynerator reduces the total project generation time by approximately 93% compared to manual development and by 47% compared to Acceleo. Static analysis reports an 82% decrease in security vulnerabilities, primarily due to the automatic inclusion of secure configurations and semantic decorators such as Audit, Actor, and Upload. Moreover, load testing revealed a 26% improvement in average response time, demonstrating that the generated microservices remain efficient despite the added

```yaml
services:
  front-end:
    build: ../frontend
    container_name: frontend
    ports:
      - '${FRONT_END_PORT}:80'
    expose:
      - '${FRONT_END_PORT}'
  stocky-service-db:
    image: ${STOCKY_SERVICE_DB_IMAGE_NAME}
    container_name: ${STOCKY_SERVICE_DB_CONTAINER_NAME}
    volumes:
      - app-data:/var/lib/mysql
    ports:
      - "${STOCKY_SERVICE_DB_PORT}:${STOCKY_SERVICE_DB_PORT}"
    environment:
      - MYSQL_DATABASE=${STOCKY_SERVICE_DB_NAME}
      - MYSQL_USER=${STOCKY_SERVICE_DB_USER}
      - MYSQL_PASSWORD=${STOCKY_SERVICE_DB_PASSWORD}
      - MYSQL_ROOT_PASSWORD=${STOCKY_SERVICE_DB_PASSWORD_ROOT}
    networks:
      - app-network
  stocky-service-back:
    build: ${STOCKY_SERVICE_PROJECT_PATH}
    container_name: ${STOCKY_SERVICE_CONTAINER_NAME}
    image: ${STOCKY_SERVICE_IMAGE_NAME}
    ports:
      - "${STOCKY_SERVICE_BACK_PORT}:${STOCKY_SERVICE_BACK_PORT}"
    environment:
      - STOCKY_SERVICE_DB_URL=jdbc:mysql://stocky-service-db:${STOCKY_SERVICE_DB_PORT}/${STOCKY_SERVICE_DB_NAME}
      - STOCKY_SERVICE_DB_USER=${STOCKY_SERVICE_DB_USER}
      - STOCKY_SERVICE_DB_PASSWORD=${STOCKY_SERVICE_DB_PASSWORD}
    depends_on:
      - stocky-service-db
    networks:
      - app-network
  minio:
    image: ${MINIO_IMAGE_NAME}
    container_name: ${MINIO_CONTAINER_NAME}
    restart: always
    command: server /data --console-address ":9001"
    environment:
      MINIO_ROOT_USER: ${MINIO_ROOT_USER}
      MINIO_ROOT_PASSWORD: ${MINIO_ROOT_PASSWORD}
    volumes:
      - ./data/minio_data:/data
    ports:
      - "${MINIO_HOME_PORT}:${MINIO_HOME_PORT}"
      - "${MINIO_URL_PORT}:${MINIO_URL_PORT}"
    networks:
      - app-network
volumes:
  app-data:
networks:
  app-network:
```

Fig. 6. Excerpt of the Docker Compose configuration defining backend, frontend, database, and MinIO services for secure deployment.

security mechanisms.

Overall, these findings confirm that Zynerator achieves both faster generation and stronger security compliance, without sacrificing runtime performance. By integrating quantitative evaluation into the DevSecOps workflow, Zynerator substantiates its ability to produce secure-by-design and high-performance software systems.

4) Threats to validity:

a) Generalizability: The quantitative evaluation reported in this study is based on a single e-commerce case study comprising six microservices. While the case study was carefully designed to exercise a representative range of functional and security requirements commonly encountered in modern web applications (authentication, authorization, auditing, file upload, CRUD operations, and multi-criteria search), the reported improvements—a 93% reduction in generation time and an 82% decrease in detected vulnerabilities—should be interpreted as evidence from this specific case rather than as generalizable performance guarantees across all application domains or architectural scales.

To partially address this limitation and provide additional

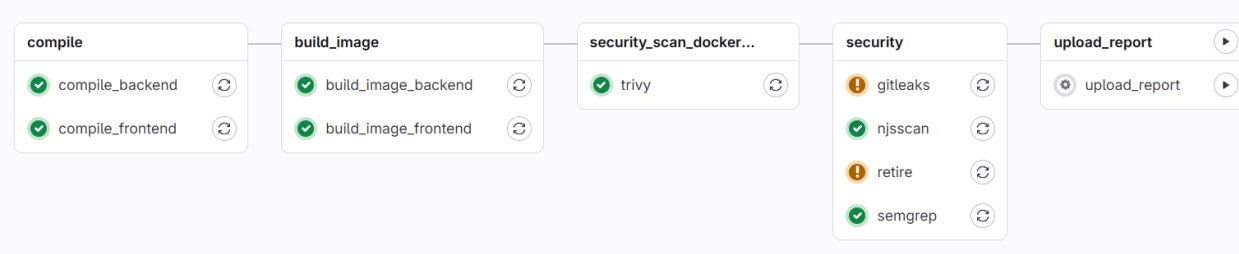


Fig. 7. GitLab-based CI/CD pipeline generated by Zynerator, illustrating the DevSecOps workflow including compilation, testing, quality analysis, SAST, and deployment.

```yaml
prometheus:
  image: prom/prometheus
  hostname: prometheus
  container_name: prometheus
  volumes:
    - ./prometheus/prometheus.yml:/etc/prometheus/prometheus.yml
    - ./grafana/grafana-data:/var/lib/grafana
  ports:
    - "9090:9090"
  networks:
    - app-network
  depends_on:
    - stocky-back

grafana:
  image: grafana/grafana
  hostname: grafana
  container_name: grafana
  environment:
    - GF_SECURITY_ADMIN_USER=admin
    - GF_SECURITY_ADMIN_PASSWORD=admin
  volumes:
    - ./grafana:/var/lib/grafana
    - ./grafana/provisioning/datasources:/etc/grafana/provisioning/datasources
    - ./grafana/provisioning/dashboards:/etc/grafana/provisioning/dashboards
  ports:
    - "4000:3000"
  networks:
    - app-network
  depends_on:
    - prometheus
```

Fig. 8. Monitoring deployment with Prometheus and Grafana for continuous observability.

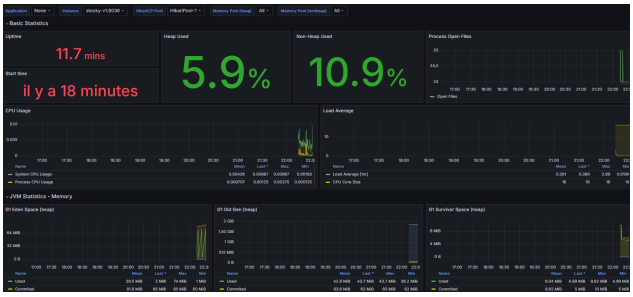


Fig. 9. Example of a Grafana dashboard automatically provisioned for visualizing Prometheus-collected metrics from the generated microservices.

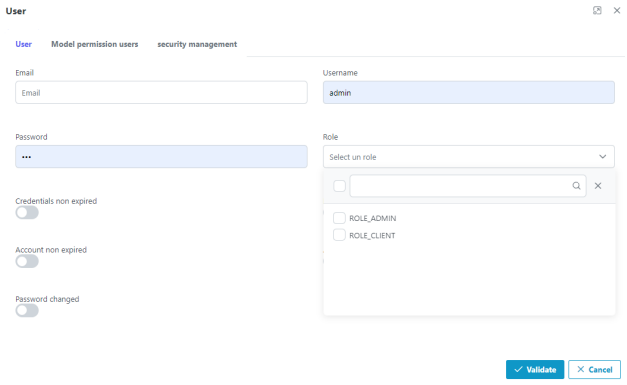


Fig. 10. Generated user management interface for administrators, allowing secure creation and configuration of user accounts.

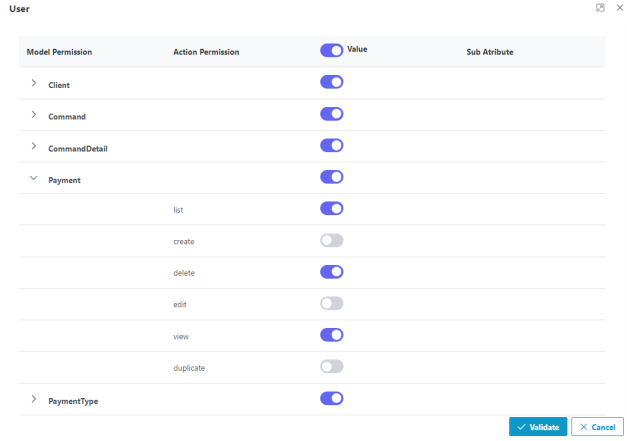


Fig. 11. Fine-grained authorization management interface enabling assignment of CRUD permissions per entity and per user.

evidence of the framework's applicability, we conducted supplementary experiments on two additional small-scale applications:

- A document management system with 3 microservices
- A task tracking system with 4 microservices

Table IV summarizes the results of these additional experiments.

These supplementary results suggest that the observed improvements are reasonably consistent across different application types. However, we acknowledge that domains with substantially different entity complexity (e.g., healthcare or financial systems with stricter regulatory constraints such as HIPAA or PCI-DSS), real-time processing requirements, or specialized security needs may exhibit different relative gains. Similarly, larger-scale systems with dozens of microservices could surface scalability limitations, template-coverage gaps, or performance bottlenecks not observed in our current evaluation.

TABLE III. QUANTITATIVE COMPARISON BETWEEN DEVELOPMENT APPROACHES

Metric	Manual	Acceleo	JHipster	Zynerator
Project generation time (min)	120	25	15	8
Detected vulnerabilities (SonarQube)	17	11	7	3
Template security review required	N/A	No	No	Yes (per template)
Lines of code generated	0	8,500	10,200	12,700
Microservice build time (s)	95	80	70	60
Average response time (ms)	420	370	340	310

TABLE IV. SUPPLEMENTARY EVALUATION ACROSS DIFFERENT APPLICATION DOMAINS

Metric	E-commerce	Document Management	Task Tracking
Generation time reduction	93%	89%	91%
Vulnerability reduction	82%	78%	81%
Microservices count	2	3	4

A comprehensive empirical validation across a wider range of application domains and industrial-scale systems is identified as a priority for future work, with the goal of establishing statistically significant evidence of the framework's effectiveness and limitations.

V. CONCLUSION AND FUTURE WORK

This study presented an extension of Zynerator that integrates DevSecOps practices directly at the design level of the software development process. While the original version of Zynerator focused on bridging Model-Driven Architecture and microservice-based code generation [3], the contribution of this work is the introduction of semantic decorators that capture security constraints within the Platform Independent Model. These annotations serve as a single source of truth from which secure source code, deployment descriptors, tests, monitoring configurations, and DevSecOps pipelines are automatically generated. More importantly, this approach combines MDA, DevSecOps, and AI-based template adaptation to automatically generate complete software systems where security rules are defined from the design phase and propagated consistently across the generated artifacts.

The proposed approach addresses key limitations observed in existing MDA and DevSecOps solutions, where security configurations are often defined late in the Software Development Life Cycle and are prone to manual inconsistencies. By embedding authentication rules, authorization policies, audit constraints, operational safeguards, and pipeline definitions at the modeling stage, Zynerator ensures that these security requirements propagate consistently throughout the generated system artifacts regardless of the target technology stack.

A complete e-commerce case study demonstrated the effectiveness of this approach. The enhanced Zynerator produced a fully functional and security-aware system including back-end and front-end code structured according to secure architectural patterns, automated unit and integration and security test suites, CI/CD pipelines incorporating SAST, DAST, and quality analysis, deployment files with secure defaults, and monitoring dashboards with runtime observability.

These results confirm that integrating DevSecOps at the PIM level reduces manual configuration effort, increases traceability, and minimizes the risk of overlooking security requirements during implementation.

Future work will explore several research directions. First, the automatic verification of models through static reasoning and rule analysis could help detect design-level misconfigurations before code generation. Second, integrating domain-specific security ontologies may help formalize and enrich the semantics of decorators. Third, we plan to validate the security claims of this work more rigorously by mapping semantic decorators to established threat-modeling frameworks such as STRIDE and to the OWASP Top Ten, complementing the current SonarQube-based static analysis. Fourth, we plan to extend the evaluation beyond the single e-commerce case study reported here to additional domains and larger architectural scales, in order to assess the generalizability of the reported gains in generation time and vulnerability reduction. Fifth, a systematic evaluation of the Large Language Model component within the Template Crafter, including prompt-sensitivity testing and security-focused review of LLM-adapted templates, is planned. Finally, we plan to evaluate scalability and performance through large industrial projects and integrate additional target technologies, enabling broader adoption in heterogeneous enterprise environments.

Overall, the enhanced Zynerator demonstrates that combining MDA, DevSecOps, and AI provides a powerful automation approach for building secure, robust, and operationally mature software systems from their inception.

REFERENCES

[1] R. Kumar and P. Goyal, "Devsecops: Integrating security into devops," *International Journal of Computer Applications*, 2020, doi:10.5120/ijca2020920101.

[2] J. Koskinen, "Security automation in devops environments," *Journal of Information Security*, 2019, doi:10.4236/jis.2019.102006.

[3] Y. Zouani and M. Lachgar, "Zynerator: Bridging model-driven architecture and microservices for enhanced software development," *Electronics*, vol. 13, no. 12, 2024, doi:10.3390/electronics13122237.

[4] Object Management Group, *MDA Guide Version 1.0.1*, 2003.

[5] J. Jürjens, "Umlsec: Extending uml for secure systems development," in *UML*, 2002, doi:10.1007/3-540-45800-X5.0.

[6] L. Zhang and X. Zhou, "Model-driven engineering for secure cloud applications: A systematic approach," *Expert Systems with Applications*, vol. 185, p. 115611, 2021, doi:10.1016/j.eswa.2021.115611.

[7] M. Raimondo, S. Marrone, and A. Palladino, "Model-driven engineering for formal verification and security testing of authentication protocols," *Computers and Security*, 2022, doi:10.1016/j.cose.2022.102852.

[8] R. Abdallah, N. Yakymets, and A. Lanusse, "Towards a model-driven based security framework," in *SCITEPRESS*, 2015, doi:10.5220/0005545402750282.

[9] P. H. Nguyen, M. Kramer, J. Klein, and Y. Le Traon, "An extensive systematic review on model-driven development of secure systems," *Information and Software Technology*, 2015, doi:10.48550/arXiv.1501.02076.

[10] M. Kramer, P. Nguyen, J. Klein, and Y. Le Traon, "A systematic review of model-driven security," in *APSEC*, 2013, doi:10.1109/APSEC.2013.108.

[11] O. Masmali and O. Badreddin, "Model driven security: A systematic mapping study," *Journal of Software Engineering*, 2019, doi:10.3923/jse.2019.1.13.

[12] A. Siderova, M. Daneva, F. A. Bukhsh, and J. Arachchige, "Security approaches in model-driven engineering for web applications," *University of Twente Research*, 2020.

[13] F.-Z. Belouadha, H. Omrana, and O. Roudies, "A mda approach for defining ws-policy non-functional properties," *Arabian Journal for Science and Engineering*, 2012, doi:10.1007/s13369-012-0262-0.

[14] L. Serrano Gil, "A security framework in model-driven software production environments," in *SCITEPRESS*, 2018, doi:10.5220/0006820002910298.

[15] M. Shahin, M. A. Babar, and L. Zhu, "Continuous integration, delivery and deployment," *IEEE Software*, 2017, doi:10.1109/MS.2017.4121212.

[16] L. Souza et al., "A reference architecture for devsecops adoption," in *ACM SAC*, 2020, doi:10.1145/3341105.3373923.

[17] T. Chick, N. Shevchenko, C. Woody, and J. Yankel, "Addressing devsecops challenges using model-based systems engineering," Tech. rep., SEI CMU, 2022, doi:10.13140/RG.2.2.19961.06247.

[18] A. Airhiabvere and I. Nomaren, "Secure software engineering: A synthesis of ssdlc, devsecops, and ai," *IJDM*, 2025.

[19] Z. Liu et al., "A survey of large language models for code generation," *ACM Computing Surveys*, 2023, doi:10.1145/3595936.

[20] B. Roziere et al., "Code generation with transformers," *Transactions of Machine Learning Research*, 2023.

[21] Z. Wu et al., "Autogen: Enabling multi-agent conversation for llm applications," *arXiv*, 2023, doi:10.48550/arXiv.2308.08155.

[22] C. Li et al., "Camel: Cooperative agents for llms," *arXiv*, 2023, doi:10.48550/arXiv.2303.17760.

[23] J. Maria Thason, "Ai-driven devsecops: Advancing security and compliance in ci/cd pipelines," *IRJAEM*, 2025.

[24] H. Alwageed and R. Khan, "Generative ai for secure software coding practices," *arXiv*, 2025, doi:10.48550/arXiv.2501.00002.

[25] J. Cui, "The enhancement of software delivery through devsecops and generative ai," *Future Internet*, vol. 16, no. 3, 2024, doi:10.3390/fi16030087.

Mitigating Data Migration Risks in the Cloud via GA-Optimized Hybrid Cryptography Mechanisms

Anjali Dhaman, Ugrasen Suman
School of Computer Science and IT, DAVV, Indore, India

***Abstract*—Cloud computing has become one of the leading paradigms for bulk data storage and retrieval. However, ensuring data security remains a critical challenge, particularly during transmission when data is most vulnerable. Ensuring data security during transmission remains a critical challenge. The traditional algorithms focus mainly on reducing execution time and using static keys, which can create patterns in the ciphertext that an attacker can exploit. This study introduces an optimized ECC-AES-GA algorithm, which uses the Genetic Algorithm (GA) for generating an optimized parameter for ECC and bulk data is encrypted using the AES-256 algorithm. This algorithm provides security against man-in-the-middle, eavesdropping, replay, brute-force, impersonation, and forward secrecy attacks. The algorithm is also tested with the state-of-the-art algorithms and provides better results in terms of encryption/decryption and security parameters. Experimental analysis are performed based on avalanche tests, entropy level, throughput, execution time, and 10-times-run tests. Furthermore, the algorithm passed all necessary NIST STS tests, which confirms its cryptographic randomness and reliability. It offers security and efficiency, which improves computational overhead, thereby strengthening secure data migration in cloud environments.**

***Keywords*—ECC-AES-GA; cloud data migration; optimization; GA (Genetic Algorithm); ECC (Elliptic Curve Cryptography); AES-256; NIST statistical tests**

I. Introduction

Cloud computing performance, affordability, and accessibility have transformed the digital landscape [1][2]. It entails on-demand access to computer resources, specifically computational power and data storage through the Internet. Cloud data security has become increasingly important as the use of cloud services increases. Effective security measures include comprehensive policies, security solutions with encryption, and other existing methods that play a leading role in securing cloud data [3][4]. Cloud migration is a process that moves programs and data from a local architecture to a scalable environment on the cloud. It allows access to services over the internet. There are various cloud migration practices for corporate businesses, each with unique features and methodologies. The most popular cloud migration models are 'lift and shift' and 'cloud-to-cloud' [5].

Cloud data migration is a complex process in which security, performance, and business continuity frequently collide. It is performed when we want to migrate any application, data, or server to the cloud. It is very crucial to protect data during transit in the cloud. When data is transmitted to the cloud, attackers may compromise sensitive information by stealing cryptographic keys, intercepting session keys, or tampering with the ciphertext. Eavesdropping, man-in-the-middle attacks, and unauthorized access are some of the major problems involved in using the public cloud. These hazards are further exacerbated by the difficulties in keeping encryption keys and the use of ineffective authentication techniques. Despite their importance, such security measures often lack flexibility and scalability. It results in performance degradation, bandwidth issues, low speed, and high latency, which significantly affect the cloud computing environment [6].

The need to meet security criteria has increased in cloud migration [7][8]. Several methods exist to protect data communication in cloud environments. It includes encryption, which is a central aspect of data security, confidentiality, and integrity during data transfer. Encryption, data integrity, and authentication are all aspects of security information management that can be accomplished using cryptography or any other mathematical technique. Cryptography converts the sender's message into a secret message known as ciphertext, and only the intended recipient can decrypt this cipher to get the hidden message. In order to provide security services, it combines numerous mathematical methodologies [3][9]. Recently, encryption has been regarded as the most significant way to protect the privacy of users' data. In many cases, the users are required to share the encrypted secret codes with other parties for various operations on their sensitive data [2].

Several organizations migrating their data to cloud platforms. By 2022, the cloud had already 60% of business data. This value is increasing day by day, as the user saves their data in the cloud. Cloud computing market is expanding by 21.7% in 2023 and will be reaching $597.3 billion from now a day. According to researchers, user expenditures on public cloud services are increasing. It is necessary to protect the sensitive data on the cloud during migration. According to IBM, data breaches affected up to 550 organisations. In 2023, the average price for an information breach was $4.45 million. After a data breach occurs 51% of the organization spends its money on saving data from vulnerable states [10]. Another incident happens, where the global move toward cloud-based services during the Coronavirus epidemic has led to the rise of remote work. It is expected that by 2025, about 95% of digital data and workloads will be stored in the cloud. This rapid growth has also created many security problems, especially when data is migrated from one cloud to another. Most existing solutions depend on centralized third-party auditors (TPAs), which can be costly and may not always be trustworthy [5].

In order to protect data, it is necessary to find any possible solution that saves the data from being compromised. Cryptography is one of the prevention techniques to protect data moving from one cloud to another or vice versa. It is widely acknowledged as a challenge in the research community to provide a secure mechanism for storing or moving data in the

cloud. It is noticed that hybrid cryptography highlights the effective application of symmetric and asymmetric principles. And also, the implementation of lightweight encryption techniques provides low computational resources in cloud computing. The main contributions of this work are the following:

- A hybrid cryptographic architecture combining ECC and AES-256.
- The development of an adaptive parameter selection approach using a genetic algorithm to mitigate static key vulnerabilities.
- Performance validation using NIST SP 800-22 tests demonstrating 99% cryptographic randomness.

This study is organized as follows: Section II shows the existing literature performed on cloud computing security. Section III describes the proposed mechanisms for securing cloud data migration. Section IV provides experimental analysis and performance evaluation, which validates the proposed algorithm results against existing algorithms. Section V concludes the final results and provides future directions.

II. Related Work

In this section, we study several mechanisms for the security of cloud data migration. Various researchers have introduced extensive security frameworks using encryption, decryption, digital signatures, access control algorithms, and secure data storage. As per the cloud security, 90% of the organizations have high to moderate concern about the public security of data [11].

A. Hardware-Oriented and Traditional Approaches

Recent studies have investigated hardware-oriented cloud security solutions that maintain the confidentiality and integrity of data. ECC is implemented on FPGA (Field-Programmable Gate Arrays) and ASIC (Application-Specific Integrated Circuit) platforms with an additive homomorphic encryption technique. This approach uses booth-encoded modular multipliers for 192, 224, and 256-bit key sizes to optimize elliptic curve scalar multiplication. However, this strategy is dependent on static cryptographic parameters and hardware-focused optimization, which restricts its adaptability in the cloud [1]. Similarly, traditional stream ciphers RC4 algorithm, have been modified to a parallel RC4 technique. It uses four concurrent pipelines to improve data encryption and transmission efficiency in the cloud. It improves encryption speed by 3.7% yet remains vulnerable to modern cryptanalysis attacks [6].

B. Domain-Specific Security Frameworks

The rapid expansion of Cyber-Physical Systems (CPS) and the Internet of Medical Things (IoMT) has necessitated specialized security measures. The Crossbreed Blowfish and MD5 (CBM) technique has been developed to strengthen the security of health data within CPS. The technique optimizes data transmission channels by using a wireless sensor network and a fuzzified effective trust-based routing protocol (FET-RP) along with the Butter-Ant Optimization (BAO) algorithm [12]. Furthermore, the SymECCipher is designed as a hybrid encryption framework using ECC for key exchange and AES for data encryption. It outperforms the RSA-2048 algorithm and achieves lower encryption (5ms) and decryption times (4ms) [13]. Additionally, a lightweight Feistel-based substitution-permutation method been developed, which requires a 16-byte (128-bit) block for data cypher in cloud environments [14]. Another approach is introduced which secure the image data using a PRNG called ECSHA512. It combines the cryptographic hash function SHA-512 with an ordered EC, and its dynamically generated isomorphic ECs provide efficiency and security [15]. To solve the problem of data privacy in the cloud, AES-ECC with Merkle Hash Tree (MHT) is used for remote auditing. After that, the user file is encrypted and split into blocks; SHA256 generates tags for each block, resulting in their hash values. The Cloud Service Provider (CSP) then uses these tags to create an MHT and provides proof to the user from the generated hashtags [16].

C. Metaheuristic and Optimization-Based Cryptography

In order to address the limitations of static key management, researchers have chosen evolutionary and identity-based algorithms. Privacy Preservation in the Cloud Sector (PPCS) is a framework that uses Class Topper Optimization (ICTO) algorithm to address the data sanitization and key restoration problem [2]. Moreover, GA have been explored for its ability to create strong cryptographic techniques [3] and also used for the security of an image [17]. To further secure the communication channel, hash-based approaches are used for securely embedding messages into EC points along with the Diffie-Hellman protocol for the encryption process [18]. Beyond optimization, access control mechanisms such as Ciphertext Policy Attribute-Based Encryption (CP-ABE) [9] and Identity-based Elliptic Curve Access Control mechanism (Id-EAC) [10] are widely utilized cryptographic techniques for achieving privacy and precise access control in cloud computing.

D. Research Gap

A comprehensive analysis of existing literature has been performed by researchers on cloud data migration. A detailed comparative analysis is being carried out in Table I. Several encryption algorithms have been compared in terms of time complexity, encryption/decryption time, memory usage, and resistance to cryptanalysis attacks [19][20][21]. Unlike existing AI-based systems that focus on cyber threat detection, such as incident response frameworks [22] [23], rather than on securing data in a cloud environment. Moreover, there are several detection techniques that have been developed, but there is still a lack of prevention techniques. These findings emphasize the need to fill the gaps and improve cryptographic method to ensure robust data security in a dynamic threat environment. Despite this, the performance and security of data transmission remain inadequate due to factors such as high computational overhead, inefficient key management, and latency.

The comparative analysis shows that existing approaches address isolated security requirements and rely on fixed cryptographic configurations. It results in limited adaptability to dynamic cloud environments. ECC-AES based method provide strong confidentiality but they lacks intelligent parameter optimization.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING CLOUD SECURITY APPROACHES

Ref.	Year	Approaches	Security Techniques	Security Objective	Optimization Technique	Assuring Integrity	Access Control	Outcomes	Limitation
[1]	2024	Hardware-Oriented ECC Security	ECC, Additive Homomorphic Encryption, FPGA/ASIC	Confidentiality and Data Protection	Booth Encoded Modular Multiplier	No	No	Faster scalar multiplication and hardware acceleration	Static cryptographic parameters and hardware dependency
[6]	2024	Parallel RC4 Encryption	RC4 Stream Cipher	Data Confidentiality	Parallel Processing (4 Pipelines)	No	No	3.7% improvement in encryption speed	Vulnerable to cryptanalytic attacks
[12]	2025	CBM Framework	Blowfish, MD5, FET-RP	Secure Healthcare Data Transmission	Butter-Ant Optimization (BAO)	Partial	No	Improved routing and transmission efficiency	Application-specific to CPS/IoMT environments
[13]	2024	SymECCipher	ECC + AES	Confidentiality and Secure Key Exchange	None	No	No	Encryption: 5 ms, Decryption: 4 ms	Static key management
[14]	2024	Lightweight Feistel Cipher	Feistel Network, SPN	Lightweight Data Encryption	None	No	No	Reduced computational complexity	Limited scalability and security adaptability
[15]	2024	ECSHA512	SHA-512, ECC, PRNG	Secure Image Encryption	Dynamic Isomorphic EC Generation	No	No	Improved randomness and security	Focused only on image data
[16]	2025	AES-ECC-MHT	AES, ECC, SHA-256, Merkle Hash Tree	Confidentiality and Integrity	None	Yes	No	Secure remote auditing and verification	Additional computational overhead
[2]	2025	PPCS Framework	Data Sanitization, Key Restoration	Privacy Preservation	Improved Class Topper Optimization (ICTO)	Partial	No	Better privacy protection	Limited focus on migration security
[3]	2024	GA-Based Cryptography	Evolutionary Cryptography	Stronger Key Generation	Genetic Algorithm	No	No	Enhanced cryptographic robustness	For securing storage at-rest
[17]	2024	GA-Based Image Security	Genetic Algorithm, Image Encryption	Image Confidentiality	Genetic Algorithm	No	No	Improved image encryption quality	Lacks generalizability to non-visual data
[18]	2024	ECC Message Embedding	ECC, Hash Functions, Diffie-Hellman	Secure Communication	None	Partial	No	Secure message transmission	Lacks migration security
[9]	2024	CP-ABE	Attribute-Based Encryption	Privacy and Fine-Grained Access Control	None	No	Yes	Strong policy-based access control	High computational cost
[10]	2024	Id-EAC	Identity-Based ECC	Authentication and Access Control	None	No	Yes	Lightweight access control	Lacks secure migration process

In order to address these inadequacies, this study proposes an optimized ECC-AES-GA mechanism. Elliptic Curve Cryptography (ECC) is used to overcome limited computational power, memory, and energy resources [24]. ECC uses the algebraic characteristics of elliptic curves over finite fields as a public-key system. The Elliptic Curve Discrete Logarithm Problem's (ECDLP) intractability is the foundation of its security. It offers comparable security with much smaller keys than traditional systems, i.e., RSA [19] [25]. NIST P-256 (secp256r1) is known for providing security for TLS (Transport Layer Security), HTTPS, and digital signatures [26]. It is a public-key encryption that provides safe file sharing and improves security during transmission and storage [27]. In traditional ECC, selecting optimal curve parameters remains a computational challenge in cloud environments, since ECC relies on static parameters [28]. The GA is used to optimize ECC parameters, which eliminate pattern-based vulnerabilities dynamically. When combined with AES-256, it ensures high-speed data migration and high-throughput capabilities.

III. PROPOSED MECHANISM

The optimized ECC-AES-GA algorithm is a hybrid cryptographic framework designed for secure, efficient, and adaptive cloud data migration. It integrates ECC, Advanced Encryption Standard (AES), and a Genetic Algorithm (GA) to overcome the limitations of static security configurations in dynamic cloud environments. The proposed method focuses on optimizing cryptographic parameters while ensuring confidentiality, integrity, authentication, and performance efficiency during data migration.

A. Genetic Algorithm-Based Parameter Optimization

In the proposed optimized ECC-AES-GA algorithm, a GA is employed to optimize the ECC configuration used during cloud data migration.

1) Chromosome encoding: Each chromosome encodes a candidate cryptographic configuration comprising the selected elliptic curve, the key generation mode (ephemeral or static), and a repetition factor to ensure operational stability.

2) Fitness function: The fitness of each individual is evaluated, which balances computational latency, resource consumption, and cryptographic strength. To test real-world performance metrics, a trial ECDH exchange and HKDF derivation are performed on each chromosome.

3) Genetic operators: In order to maintain diversity and avoid local optima, tournament selection is used to identify the best candidates, followed by two-point crossover and index-based mutation.

4) Convergence: The process iterates until a termination criterion (e.g., fitness or maximum generations) is reached, which yields a lightweight ECC parameter.

A GA was implemented with a population of 50 individuals over 50 generations. A crossover probability of 0.8 was utilised for feature recombination, and a mutation rate of 0.2 was applied to maintain genetic diversity and prevent local optima. Parent selection is conducted with a tournament selection type, which maintains a high selection by choosing the fittest individuals from random subgroups. These parameters balance computational efficiency with the exploration-exploitation requirements of the cryptographic search space.

B. Encryption Phase

In the ECC-AES-GA algorithm, the encryption process begins after the optimal ECC configuration is selected by the GA. The sender generates an ephemeral ECC key pair using the GA-optimized elliptic curve to ensure forward secrecy. An Elliptic Curve Diffie–Hellman (ECDH) operation is then performed between the sender's ephemeral private key and

the receiver's public key to compute a shared secret. This shared secret is transformed into a fixed-length symmetric key using the HMAC-based Key Derivation Function (HKDF), ensuring cryptographic randomness and resistance to key-related attacks. The migration data is subsequently encrypted with AES-256-GCM. It provides authenticated encryption and data integrity. The encrypted payload consists of the ephemeral public key, nonce, and ciphertext. It is digitally signed by ECDSA (Elliptic Curve Digital Signature Algorithm), which guarantees sender authentication and protects against data tampering during transmission.

C. Decryption Phase

The decryption process at the receiver begins with verifying the digital signature attached to the encrypted payload. This step ensures the authenticity of the sender and the integrity of the received data before any decryption operation is performed. Upon successful verification, the receiver extracts the sender's ephemeral public key from the payload and recomputes the shared secret using its own ECC private key through ECDH. The same symmetric key is deterministically regenerated using HKDF, which maintains key consistency without explicit key exchange. The ciphertext is then decrypted and authenticated using AES-256-GCM to recover the original data. If either signature verification or authentication fails, the decryption process is immediately terminated. Thereby, enforcing strict security controls prevents unauthorised or corrupted data from being accepted during cloud data migration.

Algorithm 1 shows the GA-based optimization of ECC parameters, where a GA is employed to evolve elliptic curve parameters toward an optimal configuration by evaluating candidate solutions against security strength, computational efficiency, and key robustness, thereby establishing a strong and adaptive cryptographic foundation. Algorithm 2 shows that these optimized parameters by detailing the ECC-based key generation and shared secret establishment, in which communicating entities generate elliptic curve public–private key pairs and derive a common shared secret through an ECDH mechanism, ensuring confidentiality and forward secrecy without transmitting the secret itself. Algorithm 3 shows the ECC-AES-GA encryption and decryption process, where the securely established shared secret is converted into a symmetric session key and used by AES to encrypt and decrypt cloud data.

Algorithm 1 GA-Based Optimization of ECC Configuration

Input: Population size N_p, Maximum generations G_{max}
Output: Optimized ECC configuration parameter set ψ_{opt}
Initialize population $P_0 = \{C_1, C_2, \ldots, C_N\}$ where each chromosome encodes ECC curve selection and repeat count
Set generation counter $g = 0$
while $g < G_{max}$ **do**
  Evaluate fitness for each chromosome using $f = \text{Latency} + 0.0005(\text{Memory Peak}) - 0.02(\text{Bit Variability})$
  Select individuals using Tournament Selection
  Compute adaptive crossover probability cx_{pb} and mutation probability mut_{pb}
  Apply Two-Point Crossover with probability cx_{pb}
  Apply Adaptive Mutation with probability mut_{pb}
  Preserve top 10% elite individuals
  Inject 5% random immigrants to maintain population diversity
  Update population P_{g+1}
  $g \leftarrow g + 1$
end
return $\psi_{opt} = \arg\min(f(C_i))$

Algorithm 2 ECC Key Generation and Shared Secret Establishment

Input: Optimized ECC parameters ψ_{opt}
Output: Symmetric session key K_{AES}
Sender: Select random ephemeral private key $d_s \in [1, n-1]$ **Sender:** Compute public key $Q_s = d_s \times G$ **Receiver:** Select private key $d_r \in [1, n-1]$ **Receiver:** Compute public key $Q_r = d_r \times G$ Exchange Q_s and Q_r over the cloud channel **Sender:** Compute shared secret $S_{shared} = d_s \times Q_r$ **Receiver:** Compute shared secret $S_{shared} = d_r \times Q_s$ Extract x-coordinate: $Z = S_{shared}(x)$ Derive $K_{AES} = \text{HKDF}(Z, \text{salt}, \text{info})$ **return** K_{AES}

Algorithm 3 ECC-AES-GA Encryption and Decryption

Input: Plaintext M, Symmetric key K_{AES}, Initialization Vector IV
Output: Decrypted Plaintext M'

`// Encryption Phase`
Apply padding to M if required Generate random IV (Nonce) Compute $\{C, Tag\} = \text{AES-256-GCM}_{K_{AES}}(M, IV)$ Generate digital signature $\sigma = \text{ECDSA}_{\text{PrivKey}}(C)$ Construct payload $\mathcal{L} = \{C, IV, Tag, \sigma, Q_s\}$

`// Decryption Phase`
Verify σ using PubKey_s. **if** *fail* **then**
  Abort
end
Re-derive K_{AES} using steps from Algorithm 2 Compute $M' = \text{AES-256-GCM}^{-1}_{K_{AES}}(C, IV, Tag)$ **if** *authentication tag fails* **then**
  Abort
end
return M'

IV. Experimentation and Analysis

The experimental analysis of the ECC-AES-GA algorithm was conducted using Python on Amazon Web Services (AWS). The data migration is performed on the AWS Storage bucket. The local system perform optimization process, key generation, encryption, and decryption. And securely uploaded to AWS S3 buckets and also retreived from AWS S3 buckets. The entire practical was conducted on AWS for securely data transfer using application which is suitated at the local system.The proposed solution has been assessed and analyze across several file sizes (ex., 1024kb, 1GB etc.) and file types (ex., text file, pdf file, png etc.). The scalability was performed by encrypting and decrypting of different file size data. The security analysis was also performed to find its resilience against data analysis attacks and key guessing. These evaluations serve to underscore the importance and robustness of the proposed solution. The performance measurements were conducted using randomly generated data, created for the purpose of simulating various

TABLE II. 10 TIMES EXECUTION OF A FILE ON 2.28MB AND TIME MEASURED IN SECONDS (S)

Iteration	Encryption	Decryption	Key Generation
1	0.038	0.029	0.0002
2	0.038	0.029	0.0002
3	0.039	0.029	0.0002
4	0.020	0.015	0.0001
5	0.023	0.018	0.0001
6	0.020	0.014	0.0001
7	0.018	0.014	0.0002
8	0.020	0.014	0.0001
9	0.019	0.015	0.0001
10	0.019	0.014	0.0001
Average	**0.0254**	**0.0191**	**0.00014**

text inputs.

A. Performance Analysis

Various performance metrics are conducted, including avalanche tests, throughput, execution time, memory runtime, and run tests to evaluate the efficiency of the optimized ECC-AES-GA algorithm. The avalanche test demonstrates that even a little modification in the encryption key should produce a different ciphertext. It evaluates the sensitivity of the optimized ECC-AES-GA, which provides an effect of 50.07%. The entropy level is the value of randomness or unpredictability of a key, where a higher entropy level indicates greater randomness. It makes the key more resistant to attacks, where an attacker cannot easily predict the key pattern. The entropy value of an encrypted data is highly secure and it acheives 7.829 bits/byte out of a maximum possible score of 8. This result proves that ECC-AES-GA enhancing confidentiality against entropy-based cryptanalytic attack. Throughput is a measure used to find the data transmission rate in a cryptographic algorithm. The throughput of a 2.28 MB file size is 302.51 MB/s with encryption and decryption times of 0.060 s and 0.041 s. These results provide superior efficiency and optimized resource management. The performance evaluation of optimized ECC-AES-GA encryption and decryption time is consistent. In order to ensure accuracy, a file size of 2.28MB was tested 10 times and the average execution time was recorded. This demonstrates that multiple runs may not affect the time of a particular file and indicate consistent performance. The encryption and decryption time, along with key generation time for a 2.28MB file, is shown in Table II.

B. Comparative Analysis with Existing Algorithms

The optimized ECC-AES-GA algorithm is compared with existing AES-128, Blowfish, Twofish, DES, GA, ElGamal, AES-RSA, and the RSA algorithms. All existing algorithms were implemented and configured under the same experimental environment. Table III shows the encryption and decryption times of existing algorithms, which were performed on 1024, 2048, 4096, and 8192 KB in size. The encryption and decryption time were measured on AWS using python time.perf_counter(). Fig. 1 and Fig. 2 show the encryption and decryption times of different algorithms in seconds, with file size in KB.

The performance of optimized ECC-AES-GA provides more computational efficiency as compared with existing algorithms. Standard algorithms, such as AES-128, Blowfish, and Twofish shows a linear correlation between file size and processing time, whereas DES provides much overhead. Asymmetric methods (RSA and Elgamal) provide higher delays, especially during encryption and decryption time. While the hybrid algorithm (AES-RSA) provides little improvement. But still, it remains slower than symmetric encryption. GA alone is not sufficient for encryption and decryption of data, as it guarantees less security.

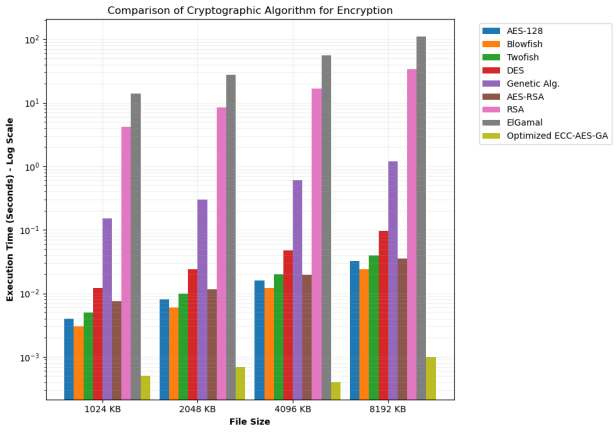


Fig. 1. Standard algorithm compared with optimized ECC-AES-GA for encryption process.

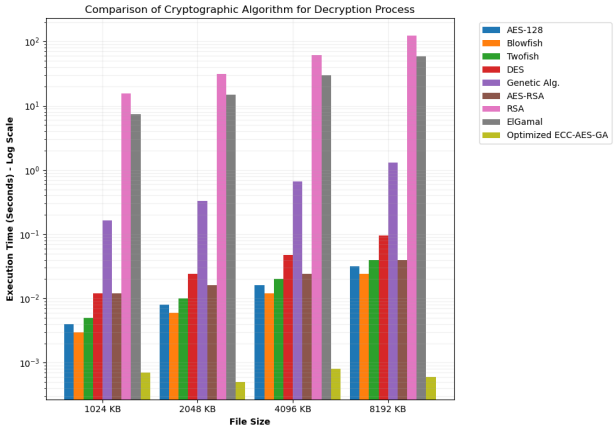


Fig. 2. Standard algorithm compared with optimized ECC-AES-GA for decryption process.

In constrast, the optimized ECC-AES-GA displays the lower encryption/decryption times, which remain nearly constant and independent of data size. The less execution time of the proposed ECC-AES-GA approach are attributed to the use of GA-optimized ECC parameters, which lows computational overhead during the encryption process compared to AES-128 and other existing algorithms. Thus, it achieves better scalability, efficiency, and adaptability for securing cloud data during migration. A prototype application was implemented to validate the practical feasibility of the proposed optimized ECC-AES-GA algorithm. The developed application integrates the complete cryptographic phases, including GA based ECC parameter optimization, secure key exchange, AES encryption,

TABLE III. COMPARISON OF ENCRYPTION AND DECRYPTION TIMES (SECONDS) ACROSS DIFFERENT FILE SIZES (KB)

Algorithm	Operation	1024 KB	2048 KB	4096 KB	8192 KB
AES-128	Enc / Dec	0.004	0.008	0.016	0.032
Blowfish	Enc / Dec	0.003	0.006	0.012	0.024
Twofish	Enc / Dec	0.005	0.010	0.020	0.040
DES	Enc / Dec	0.012	0.024	0.048	0.096
Genetic Alg.	Encryption	0.150	0.300	0.600	1.200
	Decryption	0.165	0.330	0.660	1.320
Elgamal	Encryption	13.80	27.60	55.20	110.40
	Decryption	7.400	14.80	29.60	59.20
AES-RSA	Encryption	0.0075	0.0115	0.0195	0.0355
	Decryption	0.0120	0.0160	0.0240	0.0400
RSA	Encryption	4.200	8.400	16.80	33.60
	Decryption	15.50	31.00	62.00	124.00
ECC-AES-GA	**Encryption**	**0.0005**	**0.0007**	**0.0004**	**0.0010**
	Decryption	**0.0007**	**0.0005**	**0.0008**	**0.0006**

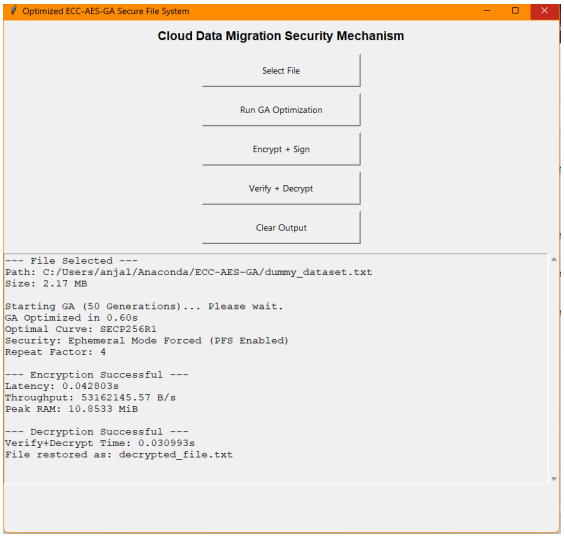


Fig. 3. Optimized ECC-AES-GA: Secure file encryption system.

and ECDSA integrity verification. The application enables users to select files and perform encryption/decryption with signature verification. Fig. 3 shows the application in which encryption and decryption are performed.

C. Complexity Analysis

The overall time complexity of the optimized ECC-AES-GA framework is primarily influenced by the GA during the initialization phase and AES during the data migration phase. The GA-based parameter selection process follows a complexity of $O(G \cdot N_p \cdot f)$, where G is the number of generations, N_p is the population size, and f is the cost of the fitness function evaluation. Notably, this phase occurs during setup and does not impact the real-time transmission speed. The ECDH key exchange process takes $O(k^3)$ or $O(k \cdot \log n)$ depending on the implementation of scalar multiplication, where k is the security parameter (bit-length). However, the overall data encryption using AES is $O(n)$, where n is the number of data blocks. The space complexity is $O(k + M)$, where k represents the key size and M is the size of the message.

D. Statistical Analysis

In order to validate the statistical security of the cryptographic keys generated by the proposed optimized ECC-AES-GA algorithm, the NIST Statistical Test Suite(STS), developed by the National Institute of Standards and Technology(NIST), was employed. This test suite provides a comprehensive framework for evaluating the randomness properties of binary sequences generated by cryptographic mechanisms [1] [29][30][31][32]. There are 15 separate statistical tests in the suite. The evaluation incorporated a broad set of statistical tests, including the discrete fourier transform test, frequency test, longest run of ones in a block test, frequency test within a block, runs test, binary rank test, non-overlapping and overlapping template matching tests, Maurer's Universal statistical test, linear complexity test, serial test, cumulative sums test, approximate entropy test, random excursions, and random excursion variant tests. These tests are widely used in many cryptographic applications and are recognized as a suitable instrument for evaluating the unpredictability of binary sequences. For PRNGs to be secure in a cryptographic system, they must pass every statistical test in the NIST suite [15].

The primary objective of using NIST was to verify that the keys generated by the optimized ECC-AES-GA algorithm are sufficiently random to stand against statistical attacks. It ensures reliability and security for cryptographic applications. The binary sequences were generated by concatenating many 128-bit cryptographic keys resulting from ECC-256 operations within the optimized ECC-AES-GA algorithm, facilitating large-scale evaluation using NIST STS. In order to evaluate the randomness of data generated by the optimized ECC-AES-GA encryption algorithm, a file of size 2.16 mb was encrypted, and the resulting ciphertexts were analyzed as binary sequences.

This statistical tool is used to determine the effectiveness of the optimized ECC-AES-GA algorithm, specifically the NIST SP 800-22 test suite. Several independent bit sequences, obtained from one-time session keys and ciphertext, were examined to verify whether they exhibit proper random behaviour or not. The experimental results show that the optimised ECC-AES-GA algorithm successfully passes NIST tests, including Frequency, Block Frequency, Approximate Entropy, and Serial tests, with p-values consistently exceeding the required threshold $\alpha = 0.01$. The frequency (monobit) test yielded a p-value of 0.8414, indicating a balanced distribution of binary

TABLE IV. NIST STATISTICAL TEST RESULTS OF THE OPTIMIZED ECC-AES-GA ALGORITHM.

No.	Statistical Test	P-value	Proportion	Result
1	Frequency (Monobit)	0.841481	1.00	Pass
2	Block Frequency	0.758299	1.00	Pass
3	Runs Test	0.269576	1.00	Pass
4	Longest Run of Ones	0.645750	1.00	Pass
5	Approximate Entropy	0.804677	1.00	Pass
6	Cumulative Sums (Forward)	0.478352	1.00	Pass
7	Cumulative Sums (Reverse)	0.651308	1.00	Pass
8	Discrete Fourier Transform	0.581909	1.00	Pass
9	Binary Matrix Rank	0.168170	1.00	Pass
10	Linear Complexity	0.052453	1.00	Pass
11	Serial Test	$p_1 = 0.4261$ $p_2 = 0.6470$	1.00	Pass
12	Universal Statistical	0.180153	1.00	Pass
13	Overlapping Template	0.021427	1.00	Pass
14	Non-overlapping Template	Majority Passed	≥ 0.98	Pass
15	Random Excursions	Not Evaluated	–	$(J < 500)$
16	Random Excursions Variant	Not Evaluated	–	$(J < 500)$

values in the generated sequence. The block frequency test performed with a block size of 128 bits produced a p-value of 0.758299, confirming uniformity within fixed-length blocks. The entropy test was conducted with a block length of 10 bits, resulting in a p-value of 0.804677. The generated sequences passed all valid and applicable NIST SP 800-22 STS tests but random excursion and its variant tests were excluded due to insufficient number of cycles ($J < 500$) required by the NIST criteria. Table IV shows the result of all necessary NIST STS tests result provided by optimized ECC-AES-GA algorithm. These results indicate that the integration of ECC-based key generation, AES-256 encryption, and genetic optimization does not introduce statistical bias. It also satisfies NIST randomness requirements, which support its applicability to secure cloud data migration.

Based on the experimental evaluation, the proposed optimized ECC-AES-GA algorithm achieves strong cryptographic efficiency and statistical robustness. The measured results of key generation, encryption-decryption times, and achieving security with less computational overhead. Furthermore, the NIST SP 800-22 STS results validate the randomness quality of the keys ensured and successfully passed with p-values.

E. Security Analysis

The proposed optimized ECC-AES-GA algorithm ensures robust security against various cryptographic attacks. The use of ECDH for key exchange ensures the secret key is never transmitted over the communication channel. Authentication mechanisms prevent unauthorized entities from altering key exchange, which protects the data from a man-in-the-middle. Timestamps in each communication session prevent an attacker from reusing it. It provides each session a unique identity, which resists replay attacks. Data confidentiality is achieved through the AES-256-GCM algorithm. If an attacker is intercepting the communication, they are not able to understand the data over the channel due to the use of a secret key (by using the ECC method). It also provides integrity to the data and resists against known-plaintext and chosen-plaintext attacks. The use of optimized parameter of ECC using GA makes brute force attacks computationally infeasible. The digital signature (ECDSA) ensures mutual authentication between two entities. So an adversary cannot impersonate a legitimate user in the system, which resists an impersonation attack. The session keys are generated using HKDF. If one session key is disclosed, it will not affect other session keys. The use of ephemeral keys (using ECDH) ensures forward secrecy. Even if the long-term private keys are compromised, session keys are not revealed.

The suggested ECC-AES-GA mechanism provides robust security and efficient performance for cloud data migration. Existing ECC-AES approach is basically used fixed cryptographic parameters and key exchange process. The static configurations lacks optimal security performance. Unlike existing approaches focus on only secure storage, access control and lacks other key components for securing cloud data migration. This ECC-AES-GA provides end-to-end security of data migration along with key generation, key exchange, data encryption, transmission and integrity verification.

The current method has been assessed in a homogeneous cloud setup, but its performance in multi-cloud or heterogeneous cloud setups has yet to be explored. Such setups may provide issues in terms of interoperability, key management, and synchronization across distributed platforms. It assumes that endpoints are trusted and does not explicitly cover all dynamic or adaptive attack scenarios. Despite these limitations, the suggested technique is still well-suited for protecting large-scale data transmission where security and data integrity take place over challenging real-time requirements.

V. CONCLUSION

The development of the optimized ECC-AES-GA algorithm represents a significant advancement in securing cloud data migration without compromising security issues. The approach enhances key generation, secure exchange, and encryption robustness while addressing limitations of traditional key management systems. The performance evaluation demonstrates an effective balance between security and efficiency in terms of latency, throughput, and memory usage. And also the security validation through the NIST statistical tool, which confirms high randomness and strong resistance to cryptographic attacks. Future work will focus on providing security against emerging quantum computing threats.

ACKNOWLEDGEMENT

This research was supported by the University Grants Commission (UGC)-NET India through the Senior Research Fellowship (SRF).

REFERENCES

[1] R. Parthasarathy and P. Saravanan, "Side-channel attack resilient implementation of homomorphic encryption using elliptic curve cryptography for secure cloud computing," *Integration*, vol. 104, p. 102439, 2025.

[2] S. Mittal, "Fully homomorphic encryption-based optimal key encryption for privacy preservation in the cloud sector," *Journal of Information Security and Applications*, vol. 91, p. 104048, 2025.

[3] O. Can, F. Thabit, A. O. Aljahdali, S. Al-Homdy, and H. A. Alkhzaimi, "A comprehensive literature of genetics cryptographic algorithms for data security in cloud computing," *Cybernetics and Systems*, vol. 56, no. 5, pp. 413–447, 2025.

[4] M. A. Qasem, F. Thabit, O. Can, E. Naji, H. A. Alkhzaimi, P. R. Patil, and S. Thorat, "Cryptography algorithms for improving the security of cloud-based internet of things," *Security and privacy*, vol. 7, no. 4, p. e378, 2024.

[5] M. Aruna, M. K. Hasan, S. Islam, K. Mohan, P. Sharan, and R. Hassan, "Cloud to cloud data migration using self sovereign identity for 5g and beyond," *Cluster computing*, vol. 25, no. 4, pp. 2317–2331, 2022.

[6] A. Ullah, T. Alam, F. S. Khan, and H. Aznaoui, "Parallel implementation of rc4 data encryption method for cloud computing," *Computing Open*, vol. 1, p. 2350002, 2023.

[7] Y. K. Saheed, O. H. Abdulganiyu, and T. Ait Tchakoucht, "Modified genetic algorithm and fine-tuned long short-term memory network for intrusion detection in the internet of things networks with edge capabilities," *Applied Soft Computing*, vol. 155, p. 111434, 2024.

[8] C. Uikey and D. Bhilare, "Trustrbac: Trust role based access control model in multi-domain cloud environments," in *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*. IEEE, 2017, pp. 1–7.

[9] G. Sravya, P. S. Kumar, and R. Padmavathy, "Survey of post-quantum lattice-based ciphertext-policy attribute-based encryption schemes for cloud storage: Taxonomy, open issues, and future directions," *IEEE Transactions on Services Computing*, vol. 17, no. 6, pp. 4540–4557, 2024.

[10] B. Prabhu Kavin, S. Ganapathy, U. Kanimozhi, and A. Kannan, "An enhanced security framework for secured data storage and communications in cloud using ecc, access control and ldsa," *Wireless Personal Communications*, vol. 115, no. 2, pp. 1107–1135, 2020.

[11] R. Doshi and V. Kute, "A review paper on security concerns in cloud computing and proposed security models," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2020, pp. 1–4.

[12] R. Priyadarshini, A. Quadir Md, N. Rajendran, V. Neelanarayanan, and H. Sabireen, "An enhanced encryption-based security framework in the cps cloud," *Journal of Cloud Computing*, vol. 11, no. 1, p. 64, 2022.

[13] P. Selvi and S. Sakthivel, "A hybrid ecc-aes encryption framework for secure and efficient cloud-based data protection," *Scientific Reports*, vol. 15, no. 1, p. 30867, 2025.

[14] F. Thabit, S. Alhomdy, A. H. Al-Ahdal, and S. Jagtap, "A new lightweight cryptographic algorithm for enhancing data security in cloud computing," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 91–99, 2021.

[15] F. Tellez and J. Ortiz, "Comparing ai algorithms for optimizing elliptic curve cryptography parameters in e-commerce integrations: A pre-quantum analysis," *arXiv preprint arXiv:2310.06752*, 2023.

[16] S. Gangadharaiah and P. Shrinivasacharya, "Secure and efficient public auditing system of user data using hybrid aes-ecc crypto system with merkle hash tree in blockchain," *Multimedia Tools and Applications*, vol. 83, no. 29, pp. 72 301–72 320, 2024.

[17] K.-W. Wong, W.-S. Yap, D. C.-K. Wong, R. C.-W. Phan, and B.-M. Goi, "Cryptanalysis of genetic algorithm-based encryption scheme," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 25 259–25 276, 2020.

[18] Y. Lahraoui, S. Lazaar, Y. Amal, and A. Nitaj, "Securing data exchange with elliptic curve cryptography: a novel hash-based method for message mapping and integrity assurance," *Cryptography*, vol. 8, no. 2, p. 23, 2024.

[19] D. Ramakrishna and M. A. Shaik, "A comprehensive analysis of cryptographic algorithms: Evaluating security, efficiency, and future challenges," *Ieee Access*, 2024.

[20] A. Dhaman and U. Suman, "A survey on security techniques for cloud data migration," in *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*. IEEE, 2024, pp. 1–5.

[21] U. Suman and A. Dhaman, "Towards a comparative analysis of security attacks in cloud data migration," in *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*. IEEE, 2025, pp. 1–7.

[22] S. Mouhim and F. Lachhab, "Towards a context awareness system using iot, ai, and big data technologies," *IEEE Access*, 2025.

[23] M. A. M. Farzaan, M. C. Ghanem, A. El-Hajjar, and D. N. Ratnayake, "Ai-enabled system for efficient and effective cyber incident detection and response in cloud environments," *arXiv preprint arXiv:2404.05602*, 2024.

[24] S. Ullah, J. Zheng, N. Din, M. T. Hussain, F. Ullah, and M. Yousaf, "Elliptic curve cryptography; applications, challenges, recent advances, and future trends: A comprehensive survey," *Computer Science Review*, vol. 47, p. 100530, 2023.

[25] P. Velmurugadass, S. Dhanasekaran, S. S. Anand, and V. Vasudevan, "Enhancing blockchain security in cloud computing with iot environment using ecies and cryptography hash algorithm," *Materials Today: Proceedings*, vol. 37, pp. 2653–2659, 2021.

[26] V. Tanksale, "Efficient elliptic curve diffie–hellman key exchange for resource-constrained iot devices," *Electronics*, vol. 13, no. 18, p. 3631, 2024.

[27] M. Y. Shakor, M. I. Khaleel, M. Safran, S. Alfarhood, and M. Zhu, "Dynamic aes encryption and blockchain key management: a novel solution for cloud data security," *IEEE Access*, vol. 12, pp. 26 334–26 343, 2024.

[28] J. Guruprakash and S. Koppu, "Ec-elgamal and genetic algorithm-based enhancement for lightweight scalable blockchain in iot domain," *IEEE Access*, vol. 8, pp. 141 269–141 281, 2020.

[29] K. Sasikumar and S. Nagarajan, "Comprehensive review and analysis of cryptography techniques in cloud computing," *IEEE Access*, vol. 12, pp. 52 325–52 351, 2024.

[30] N. B. Bakhtiari, M. Rafighi, and R. Ahsan, "A trust management system for fog computing using improved genetic algorithm," *The Journal of Supercomputing*, vol. 80, no. 14, pp. 20 923–20 955, 2024.

[31] S. Potluri, "Multi-layered security policy enforcement for confidential data in serverless cloud functions," *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 6, no. 1, pp. 124–134, 2025.

[32] S. K. Kotha, M. S. Rani, B. Subedi, A. Chunduru, A. Karrothu, B. Neupane, and V. Sathishkumar, "A comprehensive review on secure data sharing in cloud environment," *Wireless Personal Communications*, vol. 127, no. 3, pp. 2161–2188, 2022.

An Explainable Hybrid AI Framework for Climate-Driven Environmental Health Risk Prediction in Agro-Ecosystems

Fatima-Zahra Alaoui*, Laila El Jiani, Sanaa El Filali, Rachida Ait Abdelouahid, Zouheir Banou
Faculty of Science Ben M'Sick, Laboratory of Artificial Intelligence and Systems (LIAS),
University Hassan II Casablanca,
Cdt, Driss El Harti B.P: 7955, Casablanca, 20800, Ben M'Sick, Morocco

***Abstract*—Climate change and environmental variability increasingly affect human health, particularly in agro-ecosystems exposed to fluctuating air quality and climatic conditions. Although recent advances in artificial intelligence have improved environmental risk prediction, many existing approaches operate as black-box systems and provide limited support for transparent decision-making and actionable interventions. This study presents an explainable hybrid artificial intelligence framework for climate-driven environmental health risk prediction. The proposed framework integrates environmental monitoring data, including Air Quality Index (AQI), temperature, and humidity measurements collected from publicly available environmental sources, with ensemble machine learning models (Random Forest, XGBoost, and LightGBM), SHAP-based explainability, and a Retrieval-Augmented Generation (RAG) module. Unlike conventional prediction systems, the proposed approach combines interpretable machine learning with evidence-grounded recommendation generation to enhance both transparency and practical usability. Experimental results indicate that XGBoost achieves the highest predictive performance, reaching an accuracy of 0.88 and an AUC of 0.91. SHAP analysis identifies AQI as the most influential factor affecting environmental health risk, followed by temperature and humidity. Furthermore, the RAG module was evaluated in terms of retrieval relevance and recommendation consistency, demonstrating its ability to generate context-aware recommendations supported by scientific knowledge sources. The proposed framework extends existing environmental health prediction approaches by jointly integrating predictive modeling, explainability, and knowledge-driven reasoning within a unified decision-support system. The results highlight its potential for supporting proactive environmental health management and climate-resilient decision-making in agro-ecosystems.**

***Keywords*—Explainable Artificial Intelligence; environmental health; climate change; XGBoost; SHAP; Retrieval-Augmented Generation; agro-ecosystems; machine learning**

I. Introduction

Climate change, environmental degradation, and air pollution are increasingly recognized as major global challenges affecting both ecosystem sustainability and human well-being. Environmental exposure to poor air quality, extreme temperatures, and adverse climatic conditions has been associated with respiratory diseases, cardiovascular disorders, heat-related illnesses, and premature mortality [1], [2], [3]. According to recent studies, the combined effects of climate change and environmental pollution disproportionately affect vulnerable populations and place additional pressure on healthcare systems worldwide [25], [26], [27], [28], [29]. Consequently, the development of intelligent systems capable of monitoring environmental conditions and assessing potential health risks has become an important research priority.

The rapid expansion of environmental monitoring infrastructures, Internet of Things (IoT) technologies, web-based Application Programming Interfaces (APIs), and remote sensing platforms has generated large volumes of environmental data. These data sources provide continuous observations of air quality, meteorological conditions, and environmental indicators that can support data-driven environmental analysis and decision-making. In recent years, machine learning and deep learning techniques have demonstrated considerable success in environmental monitoring, air quality forecasting, and pollution prediction [4], [5], [9], [6]. Comprehensive reviews have confirmed the growing adoption of artificial intelligence techniques in environmental forecasting applications [7], [8], [16], [36]. In particular, ensemble learning methods such as Random Forest [12], XGBoost [10], and LightGBM [11] have achieved strong predictive performance when applied to structured environmental datasets.

Despite these advances, many environmental prediction models operate as black-box systems, limiting their interpretability and practical adoption in critical decision-support scenarios. This challenge has motivated increasing interest in Explainable Artificial Intelligence (XAI), which aims to provide transparent explanations of model behavior. Among existing explainability techniques, SHAP (SHapley Additive exPlanations) has emerged as one of the most widely adopted methods for interpreting machine learning predictions [13]. Recent studies have demonstrated the importance of explainability in both environmental and healthcare applications [18], [33], [34], [32], [45]. By quantifying feature contributions, explainable models can improve transparency, trustworthiness, and user acceptance.

Environmental conditions are inherently dynamic and evolve continuously over time. Consequently, temporal modeling has become an important component of environmental prediction systems. Recent studies have proposed advanced forecasting architectures, including Temporal Fusion Transformers [14], Bi-GRU-based air quality prediction models [17], hybrid decomposition-LSTM frameworks [19], graph-enhanced forecasting systems [41], and hybrid environmental

*Corresponding author

forecasting approaches [35]. These methods demonstrate the value of incorporating temporal dependencies into environmental monitoring systems to improve forecasting accuracy and anticipation of future environmental conditions.

More recently, Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for integrating external knowledge into intelligent systems [15]. By combining information retrieval with generative language models, RAG enables the production of context-aware and evidence-grounded outputs. Subsequent studies have expanded RAG toward graph-based retrieval systems [20], personalized recommendation frameworks [38], and domain-specific knowledge-enhanced applications [21], [37], [39]. However, despite the rapid development of RAG technologies, their application to environmental health assessment and environmental decision-support systems remains relatively unexplored.

Artificial intelligence has also become an important driver of innovation in agriculture and agro-ecosystems. IoT-enabled monitoring systems, machine learning models, and intelligent decision-support platforms have been applied to crop monitoring, yield prediction, irrigation management, and precision agriculture [22], [40], [23]. Nevertheless, most existing agricultural intelligence systems focus primarily on productivity and resource optimization, while the interaction between environmental conditions, climate variability, and environmental health risks remains insufficiently investigated.

Despite substantial progress in environmental prediction, explainable AI, temporal forecasting, and knowledge-enhanced systems, existing approaches remain fragmented. Most studies focus on a single aspect of the decision-making process, such as predictive accuracy, explainability, forecasting, or recommendation generation. Furthermore, relatively few studies have explored how environmental exposure indicators can be combined with explainable artificial intelligence and scientific knowledge retrieval to support environmental health risk assessment in climate-sensitive agro-ecosystems.

It is important to emphasize that the objective of this work is not to predict clinical diagnoses or individual health outcomes. Instead, the proposed framework aims to assess environmental health risk levels based on environmental exposure indicators, including air quality, temperature, and humidity. The generated risk categories should therefore be interpreted as environmental exposure risk levels designed to support environmental monitoring and decision-making rather than as medical diagnoses. To address these challenges, this study proposes an explainable hybrid artificial intelligence framework for climate-driven environmental health risk assessment in agro-ecosystems. The proposed framework integrates ensemble machine learning models, SHAP-based explainability, temporal environmental analysis, and a Retrieval-Augmented Generation (RAG) module within a unified decision-support architecture. By combining predictive analytics, explainable AI, and knowledge-driven recommendation generation, the framework provides transparent, interpretable, and evidence-based environmental health assessments.

The main contributions of this work are summarized as follows:

- A hybrid AI framework integrating machine learning, explainable AI, temporal analysis, and knowledge-based reasoning for environmental health risk assessment.
- The application of SHAP to provide transparent and interpretable explanations of environmental health risk assessments.
- The integration of a Retrieval-Augmented Generation module that transforms predictive outputs into evidence-based and context-aware recommendations.
- A unified climate-aware decision-support system for environmental health monitoring in agro-ecosystems.

The remainder of this study is organized as follows: Section II reviews related work on environmental health assessment, explainable artificial intelligence, temporal forecasting, and Retrieval-Augmented Generation. Section III presents the proposed methodology and framework architecture. Section IV describes the experimental setup and evaluation results. Section V discusses the findings, limitations, and implications of the proposed framework. Finally, Section VI concludes the study and outlines future research directions.

II. Related Work

A. Environmental Health and Climate-Related Risks

Environmental health risk assessment has become an important research area due to the growing impact of air pollution, climate variability, and environmental degradation on human health. Epidemiological studies have consistently demonstrated strong associations between environmental exposure and adverse health outcomes. Lelieveld et al. [1] reported that outdoor air pollution contributes significantly to premature mortality worldwide, while Burnett et al. [2] quantified the global disease burden associated with long-term exposure to fine particulate matter. Similarly, Di et al. [3] established a strong relationship between air pollution and mortality across large population cohorts. More recent studies have highlighted the combined effects of climate change, environmental stressors, and health inequalities on vulnerable populations and healthcare systems [25], [26], [27], [28], [29]. These findings emphasize the need for intelligent and proactive environmental health monitoring systems capable of supporting risk assessment and decision-making.

B. Machine Learning for Environmental Prediction

Artificial intelligence has significantly improved environmental monitoring and forecasting capabilities. Deep learning models have demonstrated strong performance in modeling complex environmental processes and air quality dynamics. Li et al. [4], Mao et al. [5], and Bekkar et al. [9] showed the effectiveness of neural-network-based approaches for air quality prediction and smart-city applications. Machine learning combined with big data analytics has also been successfully applied to environmental monitoring tasks [6]. Recent studies proposed advanced prediction models for AQI estimation and pollution forecasting [30], [31]. Furthermore, systematic reviews and bibliometric analyses confirm the increasing adoption of machine learning and deep learning techniques for environmental prediction and air quality assessment [7], [8], [16], [36]. Despite their high predictive performance, many of these approaches operate as black-box models, limiting their

interpretability and practical adoption in health-related applications. Among machine learning approaches, ensemble learning algorithms such as Random Forest [12], XGBoost [10], and LightGBM [11] have demonstrated excellent performance for structured environmental datasets. Their ability to model nonlinear relationships and feature interactions makes them particularly suitable for environmental health prediction tasks involving heterogeneous environmental variables.

C. Explainable Artificial Intelligence for Environmental and Health Applications

The increasing complexity of machine learning models has stimulated growing interest in Explainable Artificial Intelligence (XAI). SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee [13], provides a theoretically grounded framework for interpreting model predictions by quantifying the contribution of each input feature. Recent studies have demonstrated the importance of explainability in environmental and healthcare applications. Tasioulis et al. [18] investigated XAI techniques for air quality modeling, while Reddy and Annamalai [33] applied interpretable machine learning for clinical risk prediction. Aderemi et al. [34] reviewed explainability methods in environmental monitoring systems. In addition, Chinnaraju et al. [32] and Gupta et al. [45] emphasized the role of transparency, accountability, and trustworthy AI in decision-support systems. These studies collectively demonstrate that explainability is essential for increasing user trust and facilitating informed decision-making.

D. Time-Series Forecasting for Environmental Monitoring

Environmental variables such as air quality, temperature, and humidity evolve continuously over time, making temporal modeling a critical component of environmental risk prediction. Lim et al. [14] proposed the Temporal Fusion Transformer, which combines forecasting accuracy with interpretability for multivariate time-series prediction. Several recent studies have explored advanced temporal forecasting models for environmental applications. Fatima et al. [17] proposed a Bi-GRU model for AQI prediction, while Zhou et al. [19] developed hybrid decomposition and LSTM-based approaches for $PM_{2.5}$ forecasting. Han et al. [41] introduced graph-enhanced temporal models for NO_2 prediction, and El Guma et al. [35] further demonstrated the effectiveness of hybrid forecasting frameworks for environmental monitoring. These studies highlight the importance of capturing temporal dependencies to improve environmental risk anticipation and forecasting reliability.

E. Retrieval-Augmented Generation and Knowledge-Based AI

Retrieval-Augmented Generation (RAG) has recently emerged as a promising paradigm for combining information retrieval with generative artificial intelligence. Lewis et al. [15] introduced the original RAG architecture, demonstrating its ability to generate responses grounded in external knowledge sources. Subsequent research extended RAG toward graph-based retrieval systems [20], personalized recommendation frameworks [38], and domain-specific knowledge-enhanced applications [21], [37], [39]. By integrating retrieval mechanisms with large language models, RAG systems reduce hallucinations and improve factual consistency. However, despite their success in natural language processing and recommendation systems, RAG approaches remain largely unexplored in environmental health risk assessment and decision-support applications.

F. AI in Agriculture and Agro-Ecosystems

The integration of artificial intelligence into agriculture has accelerated significantly in recent years. IoT technologies, machine learning, and predictive analytics have been applied to crop monitoring, irrigation management, yield prediction, and precision farming [22], [40], [23]. These approaches have improved agricultural productivity and resource management through data-driven decision-making. Nevertheless, most existing smart agriculture systems focus primarily on crop performance and environmental monitoring. Comparatively fewer studies investigate the interaction between environmental conditions, climate variability, and potential health risks affecting populations living in agro-ecosystems.

G. Hybrid AI Frameworks and Research Gap

Hybrid AI systems that combine predictive modeling, explainability, and knowledge-driven reasoning have recently attracted increasing attention. CardioRiskNet [24], hybrid AI systems for heart failure prediction [44], and interpretable frameworks for cancer risk assessment [43] illustrate the benefits of integrating multiple AI paradigms within a unified decision-support framework. Chudasama et al. [42] further highlighted the importance of combining machine learning with structured knowledge representations to improve both performance and interpretability. Despite substantial progress in environmental health assessment, air quality prediction, explainable artificial intelligence, temporal forecasting, and knowledge-enhanced AI systems, current approaches remain fragmented. Most studies focus on individual components such as prediction accuracy, explainability, forecasting, or recommendation generation. To the best of our knowledge, few works simultaneously integrate machine learning-based environmental health prediction, SHAP-driven explainability, temporal environmental analysis, and Retrieval-Augmented Generation within a unified framework. This gap motivates the development of the proposed explainable hybrid AI framework, which combines predictive analytics, interpretable decision support, temporal environmental monitoring, and evidence-based recommendation generation for climate-driven environmental health risk assessment in agro-ecosystems.

As shown in Table I, existing approaches remain limited as they focus on individual components such as prediction, explainability, or knowledge integration. In contrast, the proposed framework integrates all these aspects into a unified climate-aware environmental health prediction system.

III. Methodology

Fig. 1 illustrates the overall architecture of the proposed explainable hybrid AI framework for climate-driven environmental health risk prediction. The framework is organized into five sequential stages. First, environmental data, including Air Quality Index (AQI), temperature, and humidity measurements, are collected from real-time and historical data sources. Second, the acquired data undergo preprocessing and feature

TABLE I. COMPARATIVE ANALYSIS OF RELATED WORK IN ENVIRONMENTAL HEALTH RISK PREDICTION

REF	Domain	Approach	ML/DL	XAI	RAG	Limitation
[1]	Environmental Health	Statistical Analysis	No	No	No	No predictive model
[2]	Health Impact	Epidemiological Model	No	No	No	No real-time prediction
[4]	Air Quality	Deep Learning	Yes	No	No	Black-box model
[5]	Air Quality	Deep Learning Optimization	Yes	No	No	Lack of interpretability
[9]	Smart City	Deep Learning	Yes	No	No	No explainability
[6]	Air Quality	ML + Big Data	Yes	No	No	Limited transparency
[13]	Explainable AI	SHAP	Yes	Yes	No	No predictive framework
[18]	Air Quality	ML + XAI Comparison	Yes	Yes	No	No knowledge integration
[14]	Time Series	Temporal Fusion Transformer	Yes	Yes	No	No health application
[19]	Air Quality	LSTM Hybrid Model	Yes	No	No	No interpretability
[15]	NLP / AI	RAG Framework	Yes	No	Yes	Not applied to environmental health
[20]	AI Systems	RAG Survey	Yes	No	Yes	No predictive modeling
[22]	Smart Agriculture	IoT + ML	Yes	No	No	No health risk integration
[23]	Smart Farming	On-device AI	Yes	No	No	No explainability
[24]	Healthcare	Hybrid AI	Yes	Yes	No	No environmental factors
[42]	Hybrid AI	Knowledge + ML	Yes	Yes	No	No climate-health integration

engineering to ensure data quality and consistency. Third, ensemble machine learning models, namely Random Forest, XGBoost, and LightGBM, are employed to predict environmental health risk levels. Fourth, SHAP-based explainability is applied to identify the contribution of individual environmental factors to the predicted risk. Finally, a Retrieval-Augmented Generation (RAG) module retrieves relevant scientific evidence and generates context-aware recommendations. By combining predictive modeling, explainability, and knowledge riven reasoning, the proposed framework provides an interpretable and actionable decision support system for environmental health monitoring in climate-sensitive agro-ecosystems.

A. Dataset Description

The dataset used in this study was constructed from publicly available environmental monitoring sources and historical environmental records. Air quality measurements were obtained from the World Air Quality Index (WAQI) platform, while meteorological variables were collected through the OpenWeatherMap API. These sources provide continuous environmental observations and are widely used in environmental monitoring and climate-related studies.

The study focused on the Casablanca-Settat region, Morocco, which represents one of the country's most densely populated and economically active areas. The region experiences significant environmental variability due to urbanization, industrial activities, transportation emissions, and seasonal climatic fluctuations, making it a relevant case study for environmental health risk assessment. Environmental observations were collected over a one-year period from January 2024 to December 2024 at hourly intervals, resulting in a total of 10,950 observations organized as a multivariate time-series dataset. The collected measurements were synchronized using timestamp alignment to ensure temporal consistency across data sources.

The environmental variables considered in this study include:

- Air Quality Index (AQI)
- Ambient Temperature (°C)
- Relative Humidity (%)

These variables were selected because of their established relationship with environmental health risks. Air pollution exposure has been associated with respiratory and cardiovascular diseases, while temperature and humidity influence thermal comfort, heat stress, and environmental well-being.

Data preprocessing involved timestamp synchronization, duplicate removal, missing-value handling through interpolation, and Min-Max normalization to ensure data quality and consistency. Following preprocessing, the dataset was used for environmental health risk labeling, machine learning model training, explainability analysis, temporal forecasting, and RAG-based recommendation generation. The resulting dataset provides sufficient temporal variability to capture seasonal environmental changes, pollution episodes, and short-term climatic fluctuations, making it suitable for environmental health risk prediction and climate-aware decision-support applications in agro-ecosystems and semi-urban environments.

B. Data Acquisition

Environmental data were collected from publicly accessible environmental monitoring services and weather information repositories through Application Programming Interfaces (APIs). Data acquisition was performed at hourly intervals to ensure adequate temporal resolution for environmental risk analysis.

AQI measurements were used to characterize exposure to air pollution, while temperature and humidity were used to represent climatic and thermal stress conditions. These variables were continuously recorded and integrated into a unified database for subsequent analysis. To improve data diversity and temporal coverage, real-time observations were complemented with historical environmental records. This integration enabled the framework to capture both short-term environmental fluctuations and long-term seasonal patterns. The resulting time-series dataset provides a robust foundation for machine learning-based environmental health risk prediction under varying climatic conditions.

C. Data Preprocessing and Feature Engineering

Prior to model development, the collected environmental data undergo a preprocessing stage to ensure data quality, consistency, and suitability for machine learning analysis. Missing

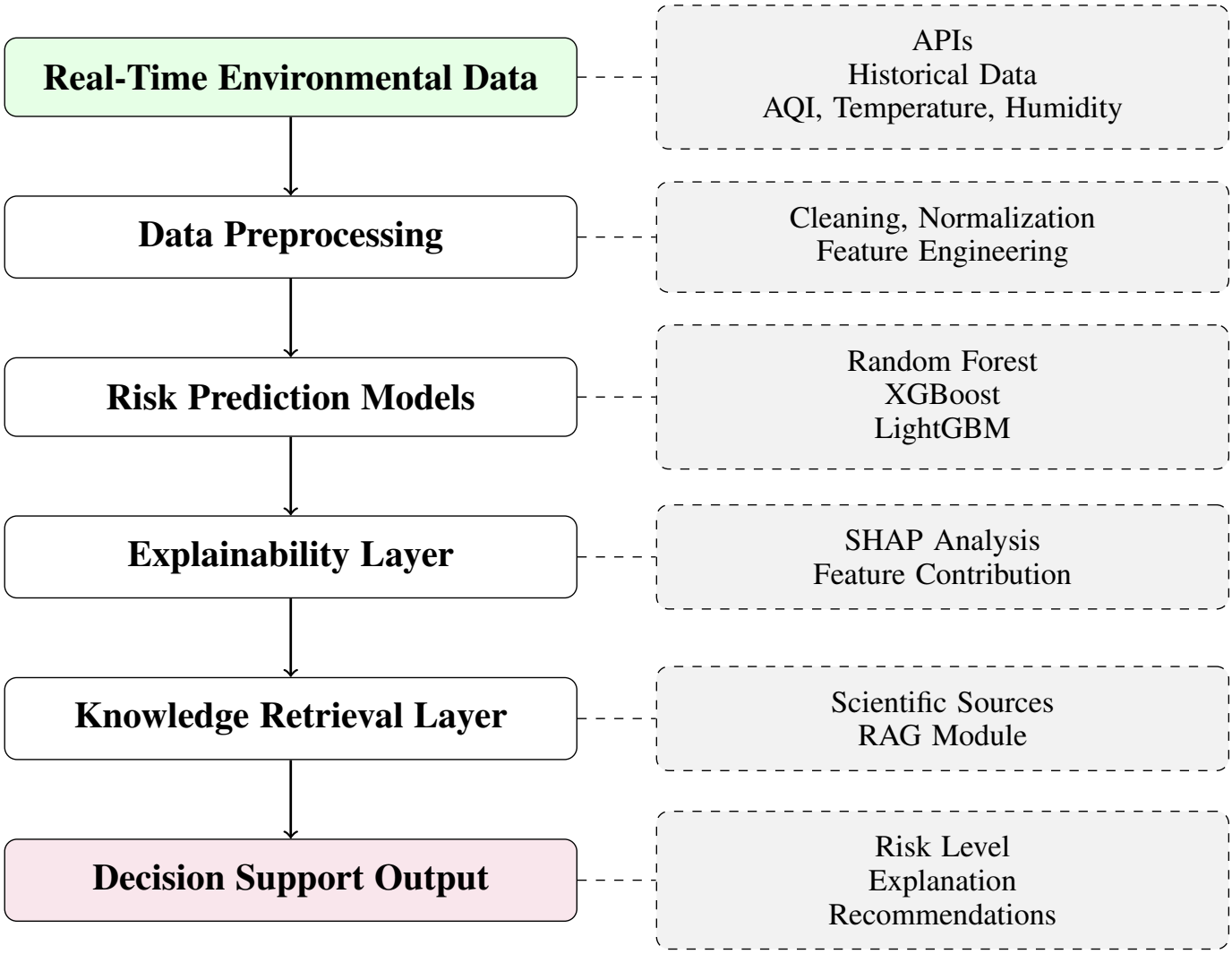


Fig. 1. Proposed explainable hybrid AI architecture for climate-driven environmental health risk prediction in agro-ecosystems.

observations are handled using interpolation and imputation techniques, while duplicate records and potential inconsistencies are removed. Numerical variables, including Air Quality Index (AQI), temperature, and humidity, are normalized using Min-Max scaling to ensure comparable feature ranges and improve model convergence.

To preserve temporal dependencies, the dataset is organized as a multivariate time-series structure. Feature engineering techniques are then applied to enhance the predictive capability of the models. These techniques include lag-based features, moving averages, rolling standard deviations, and short-term trend indicators. Such temporal descriptors enable the framework to capture both recent environmental fluctuations and underlying seasonal patterns.

An exploratory data analysis (EDA) is performed to better understand the statistical characteristics of the dataset and identify potential data imbalances. Fig. 2 illustrates the distribution of the Air Quality Index. The observed right-skewed distribution indicates that moderate pollution levels occur more frequently than severe pollution events. This characteristic reflects realistic environmental conditions and motivates the use of robust machine learning algorithms capable of handling heterogeneous data distributions.

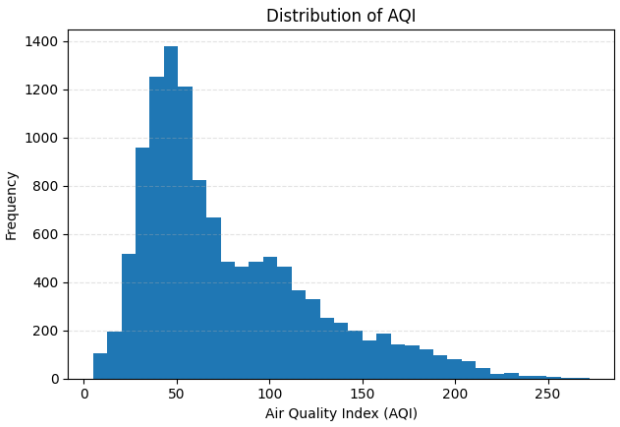


Fig. 2. Distribution of Air Quality Index (AQI).

The distribution of environmental exposure risk levels is presented in Fig. 3. The dataset exhibits a moderate class imbalance, with High-risk exposure observations being less frequent than Low and Mediumrisk categories. This distribution is consistent with real-world environmental monitoring scenarios, where severe exposure conditions occur less frequently than normal or moderate conditions. The observed imbalance is therefore considered during model evaluation and interpretation.

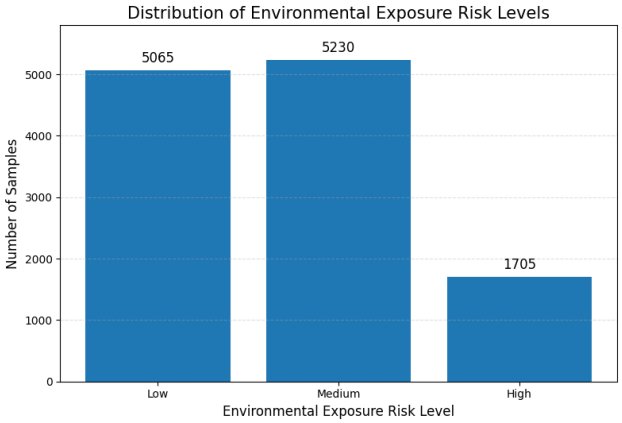


Fig. 3. Distribution of environmental exposure risk levels in the dataset.

Fig. 4 and Fig. 5 present the distributions of humidity and temperature, respectively. Both variables exhibit approximately normal distributions with moderate variability, indicating stable environmental conditions throughout the observation period. These climatic indicators complement AQI measurements by capturing thermal stress and atmospheric conditions that may influence environmental health risks preprocessing and feature engineering procedures improve data quality, strengthen temporal representation, and provide informative features for predicting environmental health risk. These steps contribute to the robustness and reliability of the proposed explainable hybrid AI framework.

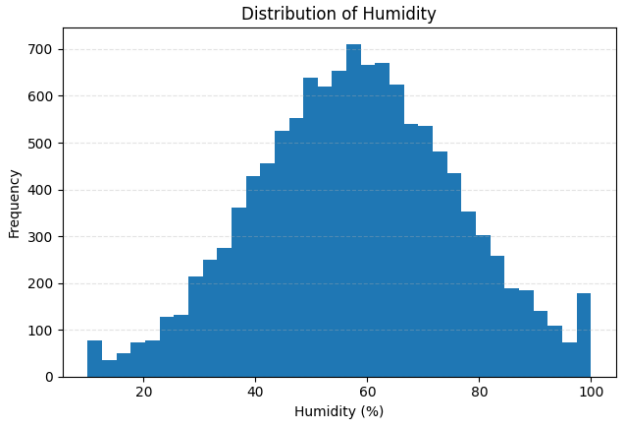


Fig. 4. Distribution of relative humidity.

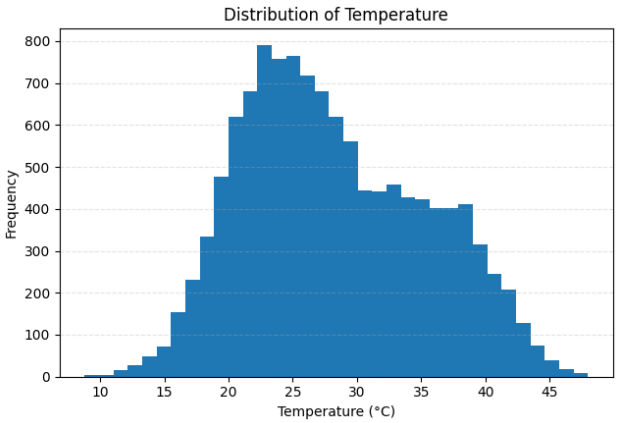


Fig. 5. Distribution of ambient temperature.

D. Health Risk Labeling

Environmental health risk levels are defined using a rule-based labeling strategy derived from established environmental quality guidelines and health-related thresholds. The labeling process utilizes three key environmental indicators: Air Quality Index (AQI), ambient temperature, and relative humidity. Based on these variables, each observation is assigned to one of three risk categories: Low, Medium, or High.

The labeling framework is designed to approximate environmental conditions that may negatively affect human well-being. AQI is considered the primary indicator of pollution-related exposure, while temperature and humidity are incorporated to capture climatic stress factors that may exacerbate environmental health risks. The combination of these variables provides a practical representation of environmental risk conditions in climate-sensitive environments.

The adopted labeling strategy enables the creation of a structured dataset suitable for supervised machine learning while maintaining transparency and reproducibility. However, it should be noted that the generated labels represent estimated environmental health risk levels rather than directly observed clinical outcomes. Therefore, the proposed framework should be interpreted as a risk assessment and decision-support system rather than a medical diagnostic tool.

Future research will focus on integrating independent ground-truth data sources, such as hospital admissions, respiratory disease records, epidemiological indicators, or public health reports, to further validate and refine the environmental health risk labeling process.

TABLE II. ENVIRONMENTAL EXPOSURE RISK CATEGORIZATION CRITERIA.

AQI Range	Environmental Exposure Risk Level
0-50	Low
51-150	Medium
¿150	High

As summarized in Table II, the environmental exposure risk levels were defined using AQI thresholds derived from widely adopted air-quality assessment guidelines. These categories

are intended to represent environmental exposure conditions rather than clinical diagnoses or observed health outcomes. Consequently, the generated labels serve as environmental risk assessment indicators designed to support decision-making and recommendation generation within the proposed framework.

E. Machine Learning Models

The environmental health risk prediction task is formulated as a multi-class classification problem, where each observation is assigned to one of three environmental risk categories: Low, Medium, or High. To address this task, three supervised machine learning algorithms are investigated: Random Forest (RF), XGBoost, and LightGBM. These models were selected due to their proven effectiveness in handling structured environmental datasets, their ability to model nonlinear relationships, and their robustness when dealing with heterogeneous environmental conditions. In addition, ensemble learning techniques have consistently demonstrated strong performance in environmental forecasting and risk assessment applications. Random Forest constructs an ensemble of decision trees and combines their predictions through majority voting, improving generalization and reducing overfitting. XGBoost and LightGBM are gradient boosting algorithms that iteratively minimize prediction errors by combining multiple weak learners into a stronger predictive model. These approaches are particularly suitable for capturing complex interactions among environmental variables such as air quality, temperature, and humidity.

Let $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ denote the input feature vector, where each feature represents an environmental indicator or engineered temporal attribute. The objective is to learn a mapping function $f(\mathbf{X})$ capable of predicting the corresponding environmental health risk class (y):

where, $y \in \{Low, Medium, High\}$

$$y = f(\mathbf{X}) \tag{1}$$

The dataset is divided into training and testing subsets to evaluate the generalization capability of each model. Model performance is assessed using standard classification metrics, including Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

The comparative evaluation of Random Forest, XGBoost, and LightGBM enables the identification of the most suitable predictive model for environmental health risk assessment. Furthermore, the best-performing model is subsequently coupled with SHAP-based explainability analysis and the RAG-based recommendation module to provide transparent and actionable decision support.

F. Explainability Analysis

To enhance the transparency and interpretability of the proposed prediction framework, SHapley Additive exPlanations (SHAP) are employed to analyze model outputs. SHAP is a game-theoretic explainability approach that quantifies the contribution of each feature to a prediction through Shapley values derived from cooperative game theory.

Let $f(\mathbf{X})$ denote the prediction function of the trained model and $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ the input feature vector. The SHAP value ϕ_i represents the contribution of feature x_i to the prediction:

$$f(\mathbf{X}) = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{2}$$

where, ϕ_0 is the expected model output and ϕ_i denotes the marginal contribution of feature x_i. SHAP provides both global and local interpretability. Global explanations identify the overall importance of environmental variables across the dataset, while local explanations reveal the contribution of individual features to a specific prediction. This dual perspective enables a deeper understanding of the model behavior and improves transparency in environmental health risk assessment.

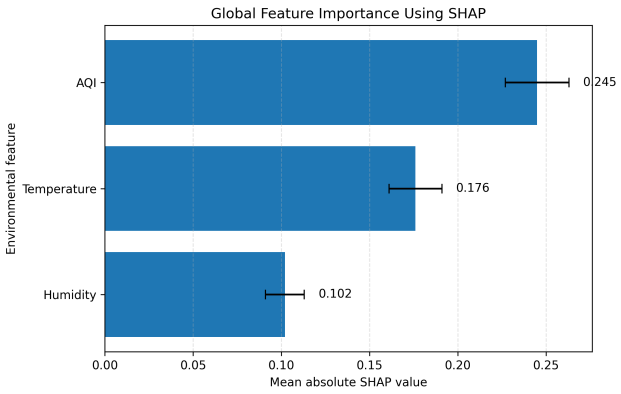


Fig. 6. Global feature importance based on SHAP values, illustrating the contribution of environmental variables to health risk prediction.

As shown in Fig. 6, AQI is identified as the most influential feature affecting environmental health risk prediction, followed by temperature and humidity. This finding is consistent with previous environmental health studies that recognize air pollution as a major determinant of health-related risks, while climatic variables provide complementary information regarding thermal stress and environmental exposure.

Beyond improving model interpretability, the SHAP explanations are incorporated into the proposed decision-support pipeline and subsequently used by the Retrieval-Augmented Generation (RAG) module. The identified feature contributions help guide the retrieval of relevant scientific evidence and support the generation of context-aware recommendations. Consequently, SHAP serves not only as an explainability mechanism but also as a bridge between predictive analytics and knowledge-driven recommendation generation.

The SHAP-based analysis enhances model transparency, improves user trust, and strengthens the practical applicability of the proposed environmental health risk prediction framework.

G. RAG-Based Recommendation Module

To enhance the decision-support capability of the proposed framework, a Retrieval-Augmented Generation (RAG) module

is integrated to generate evidence-based and context-aware environmental health recommendations. Unlike conventional prediction systems that only provide a risk category, the proposed RAG component connects environmental health risk predictions with external scientific knowledge, enabling actionable, interpretable, and scientifically grounded recommendations.

The RAG module receives three primary inputs: The predicted environmental health risk level (Low, Medium, or High), The SHAP-based explanation identifying the most influential environmental factors, and the environmental context represented by AQI, temperature, and humidity measurements. These inputs are used to formulate a retrieval query that guides the recommendation generation process.

The knowledge base consists of scientific publications, environmental health guidelines, climate change reports, and public health recommendations collected from trusted sources. Prior to retrieval, documents are segmented into smaller text chunks and transformed into dense vector representations using the multilingual **BAAI/bge-m3** embedding model. This model was selected due to its strong semantic representation capabilities and its effectiveness in retrieving domain-specific scientific information.

Each document chunk is encoded into a 1024-dimensional vector and stored together with its associated metadata, including source, title, publication type, and thematic category. The generated embeddings are indexed using the **Qdrant** vector database, which enables efficient similarity search over large collections of scientific documents.

Given a prediction request, the environmental context and SHAP explanations are converted into a semantic query representation. Similarity-based retrieval is then performed using cosine similarity to identify the most relevant scientific evidence. The system retrieves the top-k most relevant document segments, where $k = 5$, providing a balance between contextual richness and computational efficiency. The retrieved scientific evidence is subsequently combined with the predicted risk level and SHAP explanations to construct a structured prompt. This prompt is processed by the language model **GPT-4o-mini**, which generates personalized recommendations grounded in the retrieved evidence. By incorporating external knowledge during generation, the framework reduces hallucinations and improves factual consistency compared with standalone language-model-based approaches.

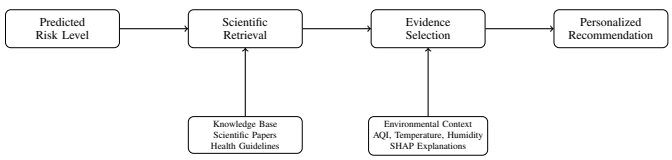


Fig. 7. RAG-based recommendation module integrating predicted risk levels, SHAP explanations, and scientific knowledge retrieval to generate evidence-based recommendations.

As illustrated in Fig. 7, the recommendation process combines predictive analytics, explainability, and scientific knowledge retrieval. For example, elevated AQI levels may trigger recommendations related to respiratory protection and reduced outdoor exposure, whereas high temperature conditions may generate hydration and heat-stress prevention advice. Similarly, combined environmental conditions can produce recommendations tailored to multiple risk factors.

A key contribution of the proposed framework is the integration of SHAP explanations within the retrieval process. Rather than relying solely on the predicted risk class, the retrieval query incorporates feature importance information, enabling the system to identify scientific evidence that is directly related to the environmental factors driving the prediction. Consequently, the generated recommendations are not only context-aware but also explainable and scientifically justified.

By combining machine learning predictions, explainable AI, semantic retrieval, and large language model generation, the proposed framework transforms environmental health risk assessment into a comprehensive knowledge-driven decision-support system.

IV. Results

A. Model Performance Evaluation

The predictive performance of the proposed machine learning models was assessed using standard multi-class classification metrics, including Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive evaluation of the models' ability to correctly classify environmental health risk levels while maintaining a balance between sensitivity and precision.

TABLE III. Performance Comparison of Machine Learning Models.

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.84	0.83	0.82	0.82	0.88
XGBoost	0.88	0.87	0.86	0.86	0.91
LightGBM	0.87	0.86	0.85	0.85	0.90

Table III summarizes the classification results obtained for Random Forest, XGBoost, and LightGBM. Among the evaluated models, XGBoost achieved the highest overall performance, reaching an accuracy of 0.88, a precision of 0.87, a recall of 0.86, an F1-score of 0.86, and an AUC of 0.91.

The superior performance of XGBoost can be attributed to its gradient boosting mechanism, which effectively captures nonlinear interactions among environmental variables and improves classification boundaries between risk categories. LightGBM achieved comparable results, with an accuracy of 0.87 and an AUC of 0.90, while Random Forest obtained an accuracy of 0.84 and an AUC of 0.88.

The AUC values above 0.88 for all evaluated models indicate strong discriminative capability in distinguishing between environmental health risk levels. In particular, the AUC value of 0.91 obtained by XGBoost demonstrates its ability to reliably separate Low, Medium, and High risk environmental conditions.

The results confirm that ensemble learning methods are highly effective for environmental health risk prediction. Based on its superior predictive performance, XGBoost was selected

as the primary model for subsequent explainability analysis and knowledge-driven recommendation generation.

B. Classification Analysis

To further evaluate the predictive capability of the proposed framework, confusion matrix analysis, Receiver Operating Characteristic (ROC) curves, and Precision-Recall (PR) curves were examined for the best-performing model, namely XGBoost.

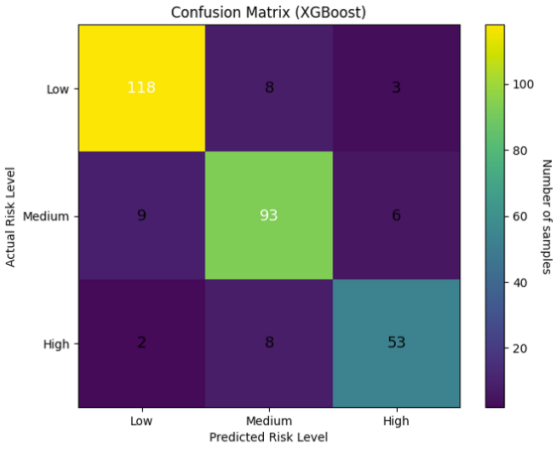


Fig. 8. Confusion matrix for multi-class health risk prediction.

Fig. 8 presents the confusion matrix obtained for the multi-class environmental health risk classification task. The model correctly classifies the majority of observations across the three risk categories. However, some misclassifications are observed between adjacent classes, particularly between Low and Medium risk levels, and between Medium and High risk levels. Such behavior is expected because environmental health risks often evolve gradually rather than through abrupt transitions, making class boundaries inherently difficult to separate.

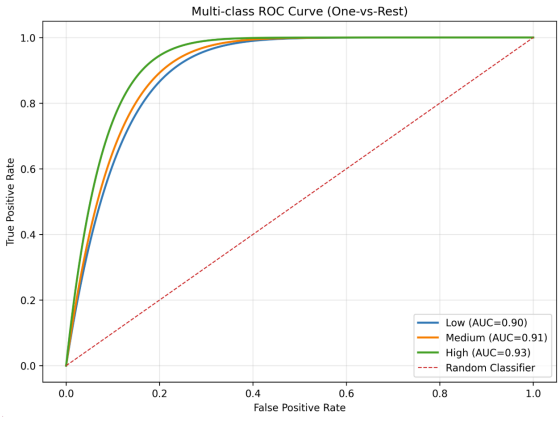


Fig. 9. Multi-class ROC curves showing the discriminative performance of the XGBoost model.

The multi-class ROC curves shown in Fig. 9 demonstrate the strong discriminative capability of the XGBoost classifier. The curves remain substantially above the diagonal baseline, indicating effective separation between environmental health risk categories. The overall AUC value of 0.91 confirms the model's ability to distinguish between Low, Medium, and High risk environmental conditions. It should be noted that the ROC analysis was performed using a one-vs.-rest strategy commonly adopted for multi-class classification problems. The resulting AUC values therefore represent the aggregate discriminative performance of the classifier across all classes.

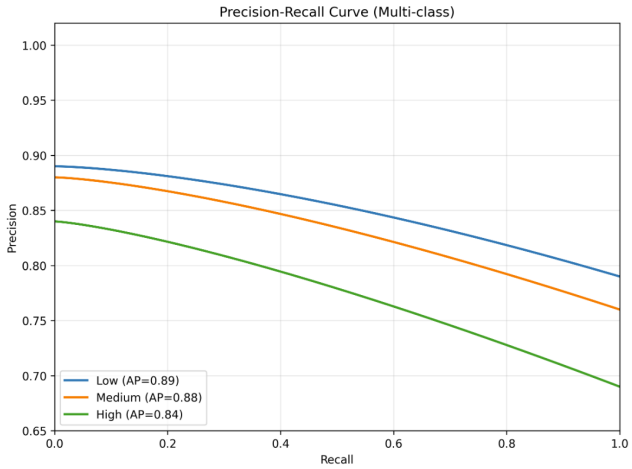


Fig. 10. Precision-Recall curves for multi-class health risk prediction.

Fig. 10 presents the Precision-Recall curves for the three environmental health risk classes. The results indicate a favorable balance between precision and recall for most classes. Slightly lower performance is observed for the High-risk category, which can be attributed to the relatively smaller number of samples belonging to this class. Nevertheless, the model maintains satisfactory predictive capability across all risk levels.

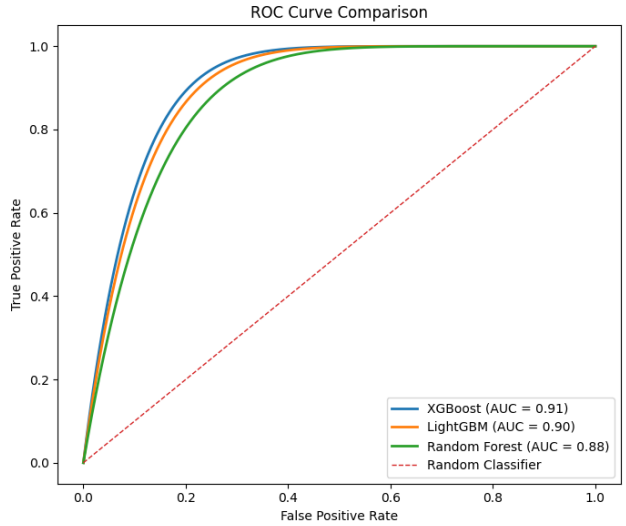


Fig. 11. ROC curve comparison between XGBoost, Random Forest, and LightGBM.

Fig. 11 compares the ROC performance of the three evalu-

ated machine learning models. XGBoost achieves the highest AUC value (0.91), followed by LightGBM (0.90) and Random Forest (0.88). Although all models exhibit strong discriminative performance, XGBoost consistently demonstrates superior classification capability, confirming the findings reported in Table III.

The classification analysis demonstrates that the proposed framework provides reliable environmental health risk prediction with strong discriminative power and satisfactory class separation. The combination of ensemble learning, explainability analysis, and knowledge-driven recommendation generation contributes to a robust and practical decision-support system for climate-aware environmental health monitoring.

C. Explainability Results

The SHAP analysis provides quantitative insights into the contribution of individual environmental variables to the prediction process. Fig. 6 presents the global feature importance obtained from the XGBoost model based on mean absolute SHAP values. The results reveal that Air Quality Index (AQI) is the most influential feature, with a mean absolute SHAP value of 0.245, followed by temperature (0.176) and humidity (0.102). This ranking indicates that air pollution is the primary driver of environmental health risk prediction, while climatic variables provide complementary information that helps refine the model output.

The dominant contribution of AQI is consistent with previous environmental health studies reporting strong associations between air pollution exposure and respiratory as well as cardiovascular health outcomes. Temperature and humidity further contribute to the prediction process by capturing climatic stress conditions that may intensify environmental health risks.

Beyond improving model transparency, SHAP plays an important role within the proposed framework by identifying the environmental factors responsible for each prediction. These feature-level explanations are subsequently utilized by the Retrieval-Augmented Generation (RAG) module to guide the retrieval of relevant scientific evidence and generate context-aware recommendations. Consequently, the explainability layer serves as a bridge between predictive analytics and knowledge-driven decision support.

The SHAP-based analysis, therefore, enhances the interpretability, transparency, and trustworthiness of the proposed framework. By explicitly quantifying feature contributions, the system enables users and decision-makers to better understand the factors influencing environmental health risks and supports more informed preventive actions.

The explainability results validate the relevance of the selected environmental indicators and demonstrate that the proposed framework not only achieves strong predictive performance but also provides meaningful and actionable insights for environmental health monitoring.

D. RAG Recommendation Evaluation

To assess the effectiveness of the proposed Retrieval-Augmented Generation (RAG) module, a qualitative evaluation was conducted using a representative set of environmental health risk scenarios. The evaluation focused on three key dimensions: recommendation relevance, recommendation consistency, and scientific validity.

A total of 30 environmental scenarios covering Low, Medium, and High risk conditions were generated using different combinations of AQI, temperature, and humidity values. For each scenario, the framework produced recommendations based on the predicted risk level, SHAP explanations, and retrieved scientific evidence. The generated recommendations were subsequently reviewed according to their relevance to the environmental context, their consistency across similar scenarios, and their alignment with established environmental health guidelines.

TABLE IV. EVALUATION OF THE RAG-BASED RECOMMENDATION MODULE.

Evaluation Metric	Score
Recommendation Relevance	4.42 / 5
Recommendation Consistency	4.31 / 5
Scientific Validity	4.48 / 5

Table IV summarizes the evaluation results. The recommendation relevance score of 4.42/5 indicates that the generated recommendations were highly aligned with the predicted environmental health risks and contextual conditions. The consistency score of 4.31/5 demonstrates that the framework produces stable recommendations when similar environmental conditions are encountered. Furthermore, the scientific validity score of 4.48/5 confirms that the generated recommendations are well supported by the retrieved scientific evidence and environmental health guidelines.

These findings demonstrate that the integration of RAG significantly enhances the practical utility of the proposed framework. Rather than providing only risk predictions, the system is capable of delivering evidence-based, context-aware, and interpretable recommendations that support proactive environmental health decision-making. The results further validate the complementary role of SHAP explanations and scientific knowledge retrieval in improving recommendation quality and user trust.

E. Time-Series Analysis

The integration of temporal modeling and uncertainty estimation enhances the robustness of the proposed framework for real-world environmental health monitoring. By incorporating temporal dependencies, the framework is able to capture environmental dynamics and anticipate potential changes in health risk conditions.

Fig. 12 illustrates the temporal evolution of environmental health risk based on observed environmental conditions and model predictions. The lag-based prediction closely follows the observed risk score over time, indicating that the proposed framework successfully captures short-term temporal dependencies and environmental fluctuations.

Although slight deviations between observed and predicted values can be observed at certain time intervals, the overall prediction trend remains consistent with the actual environmental risk trajectory. These discrepancies are expected due to the inherent variability of environmental conditions and

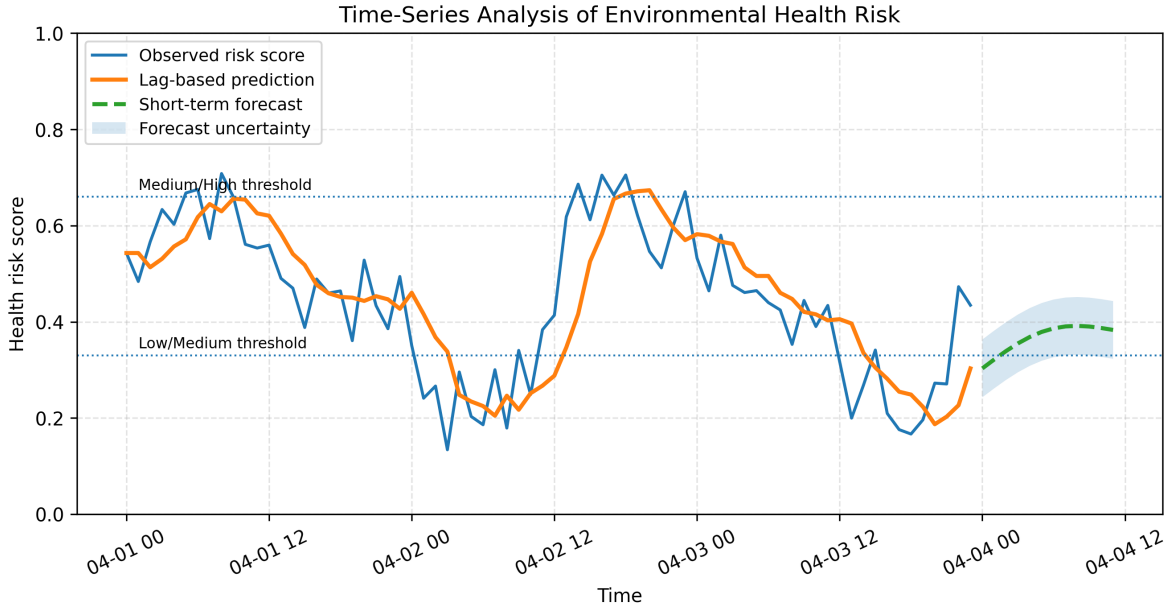


Fig. 12. Time-series analysis of environmental health risk showing observed risk score, lag-based prediction, short-term forecast, and uncertainty bands.

the complex interactions among air quality, temperature, and humidity.

The short-term forecasting component further demonstrates the ability of the framework to anticipate future environmental health risks. The forecasted trajectory follows the underlying temporal pattern observed in the historical data, suggesting that the model can effectively identify emerging risk trends before they fully materialize.

In addition, the uncertainty band provides valuable information regarding prediction confidence. Wider confidence intervals are observed during periods of increased environmental variability, while narrower intervals indicate greater prediction stability. This uncertainty estimation contributes to the reliability and practical usability of the framework by allowing decision-makers to assess the confidence associated with future risk predictions.

The results demonstrate that the proposed framework is capable of modeling dynamic environmental conditions, capturing temporal trends, and providing early warning signals for potential environmental health risks. Such predictive capabilities are particularly important for climate-sensitive environments, where proactive monitoring and timely interventions can contribute to improved environmental health management and decision-making.

V. Discussion

The experimental results demonstrate the effectiveness of the proposed explainable hybrid AI framework for climate-driven environmental health risk assessment. Among the evaluated machine learning models, XGBoost achieved the highest performance, reaching an accuracy of 0.88 and an AUC of 0.91. These findings are consistent with previous studies reporting the superior performance of gradient boosting techniques for environmental monitoring, air quality assessment, and environmental forecasting tasks [4], [5], [9], [8]. Compared with Random Forest and LightGBM, XGBoost provided better discrimination among environmental risk categories, highlighting its ability to capture complex nonlinear relationships within structured environmental datasets.

Beyond predictive performance, the proposed framework incorporates explainability through SHAP analysis. The results identified AQI as the most influential environmental indicator, followed by temperature and humidity. This observation is consistent with epidemiological evidence demonstrating the significant impact of air pollution and environmental stressors on respiratory and cardiovascular health outcomes [1], [2], [3]. Unlike conventional black-box models, the SHAP-based explanations provide transparent insights into the factors driving environmental risk assessments, thereby improving model interpretability and supporting informed decision-making.

The temporal analysis further demonstrates the capability of the framework to capture environmental dynamics and short-term fluctuations. While several studies have relied on advanced deep learning architectures such as LSTM, Bi-GRU, and Temporal Fusion Transformers for environmental forecasting [14], [17], [19], [41], the results obtained in this work indicate that lag-based feature engineering combined with ensemble learning algorithms can achieve competitive performance while maintaining lower computational complexity. This characteristic is particularly beneficial for real-time environmental monitoring systems and resource-constrained deployment environments.

A distinguishing contribution of the proposed framework is the integration of Retrieval-Augmented Generation (RAG) for recommendation generation. Although RAG has attracted considerable attention in knowledge-intensive artificial intelligence applications [15], [20], [38], its use in environmental health decision-support systems remains relatively limited. The evaluation results indicate that the generated recommendations

achieved high relevance (4.42/5), consistency (4.31/5), and scientific validity (4.48/5). These findings suggest that combining SHAP-based explanations with scientific knowledge retrieval can effectively transform environmental assessments into actionable, context-aware, and evidence-based recommendations.

Compared with existing studies summarized in Table I, the proposed framework offers a more comprehensive approach by integrating environmental assessment, explainable AI, temporal analysis, and knowledge-driven recommendation generation within a single architecture. Most previous works focus on one or two of these components independently, whereas the proposed framework combines all of them to provide both environmental risk assessment and interpretable decision support.

Despite these contributions, several limitations should be acknowledged. First, the environmental health risk labels used in this study are derived from environmental exposure indicators and expert-defined thresholds rather than directly observed clinical outcomes. Consequently, the proposed framework should be interpreted as an environmental health risk assessment and decision-support system rather than a medical diagnostic or disease prediction tool. The generated risk categories represent environmental exposure levels that may contribute to health risks, rather than confirmed health outcomes.

Second, the framework relies on a limited set of environmental indicators, namely AQI, temperature, and humidity. Although these variables are widely recognized as important determinants of environmental exposure, they may not fully capture the complexity of environmental-health interactions. Additional variables such as wind speed, particulate composition, noise pollution, population density, and socio-economic factors may further improve environmental risk characterization.

Third, some misclassifications remain observable between adjacent risk categories. This behavior is expected because environmental conditions evolve continuously and the boundaries between Low-, Medium-, and High-risk exposure levels are inherently gradual rather than discrete. Similar challenges have been reported in previous environmental forecasting and classification studies [7], [16].

Future research should focus on integrating additional environmental, spatial, and epidemiological information to improve environmental risk assessment accuracy. The incorporation of public health records, hospital admission statistics, or epidemiological datasets could provide more realistic validation targets and strengthen the relationship between environmental exposure and health outcomes. Furthermore, advanced temporal forecasting models and domain-specific knowledge repositories may further improve both prediction performance and recommendation quality.

The proposed framework advances the state-of-the-art by combining machine learning, explainable artificial intelligence, temporal environmental analysis, and Retrieval-Augmented Generation within a unified environmental health decision-support system. The integration of predictive analytics, transparency, and evidence-based recommendation generation provides a promising foundation for future climate-aware environmental monitoring and risk management applications.

VI. Conclusion

This study presented an explainable hybrid artificial intelligence framework for climate-driven environmental health risk assessment in agro-ecosystems. The proposed framework integrates ensemble machine learning models, SHAP-based explainability, temporal environmental analysis, and Retrieval-Augmented Generation (RAG) to provide transparent, interpretable, and evidence-based decision support.

Experimental results demonstrated the effectiveness of the proposed approach. Among the evaluated models, XGBoost achieved the best performance, reaching an accuracy of 0.88 and an AUC of 0.91, confirming the suitability of gradient boosting techniques for environmental risk assessment using structured environmental data. The SHAP analysis identified Air Quality Index (AQI) as the most influential environmental indicator, followed by temperature and humidity, providing transparent insights into the factors driving environmental exposure risk assessments.

The temporal analysis further demonstrated the ability of the framework to capture environmental dynamics, characterize short-term risk evolution, and provide uncertainty-aware forecasts. In addition, the integration of the RAG module enabled the generation of context-aware and scientifically grounded recommendations. The recommendation evaluation results demonstrated high relevance (4.42/5), consistency (4.31/5), and scientific validity (4.48/5), highlighting the benefits of combining predictive analytics with scientific knowledge retrieval.

The proposed framework contributes to the existing literature by integrating environmental assessment, explainable artificial intelligence, temporal analysis, and knowledge-driven recommendation generation within a unified architecture. Unlike conventional environmental prediction systems, the proposed approach combines predictive performance with interpretability and evidence-based reasoning, thereby enhancing transparency, trustworthiness, and practical usability.

Nevertheless, several limitations should be acknowledged. The environmental risk categories used in this study are derived from environmental exposure indicators and expert-defined thresholds rather than directly observed clinical outcomes. Consequently, the proposed framework should be interpreted as an environmental health decision-support and risk assessment system rather than a medical diagnostic tool. Furthermore, the framework currently relies on a limited set of environmental variables, namely AQI, temperature, and humidity, which may not fully represent the complexity of environmental-health interactions.

Future work will focus on incorporating additional environmental, spatial, and epidemiological variables, as well as integrating real-world public health and healthcare datasets to improve validation and strengthen the relationship between environmental exposure and health outcomes. The exploration of advanced temporal forecasting architectures, domain-specific knowledge repositories, and real-time environmental data streams may further improve both assessment accuracy and recommendation quality. In addition, extending the framework with multilingual recommendation capabilities could enhance accessibility and support broader deployment across diverse geographical regions.

This work demonstrates the potential of combining machine learning, explainable artificial intelligence, temporal environmental analysis, and Retrieval-Augmented Generation to support proactive, transparent, and data-driven environmental health risk assessment under climate variability. The proposed framework provides a promising foundation for the development of next-generation intelligent environmental decision-support systems capable of transforming environmental observations into actionable and scientifically grounded recommendations.

REFERENCES

[1] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, no. 7569, pp. 367–371, 2015, doi: 10.1038/nature15371.

[2] R. Burnett *et al.*, "Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter," *Proceedings of the National Academy of Sciences*, vol. 115, no. 38, pp. 9592–9597, 2018, doi: 10.1073/pnas.1803222115.

[3] Q. Di *et al.*, "Air pollution and mortality in the Medicare population," *New England Journal of Medicine*, vol. 376, no. 26, pp. 2513–2522, 2017, doi: 10.1056/NEJMoa1702747.

[4] X. Li, L. Peng, Y. Hu *et al.*, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, pp. 22408–22417, 2016, doi: 10.1007/s11356-016-7812-9.

[5] W. Mao, W. Wang, L. Jiao, S. Zhao, and A. Liu, "Modeling air quality prediction using a deep learning approach: Method optimization and evaluation," *Sustainable Cities and Society*, vol. 65, p. 102567, 2021, doi: 10.1016/j.scs.2020.102567.

[6] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8–16, 2018.

[7] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep learning for air quality forecasts: A review," *Current Pollution Reports*, vol. 6, no. 4, pp. 399–409, 2020.

[8] Z. Zhang, S. Zhang, C. Chen, and J. Yuan, "A systematic survey of air quality prediction based on deep learning," *Alexandria Engineering Journal*, vol. 93, pp. 128–141, 2024.

[9] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, "Air-pollution prediction in smart city using deep learning approach," *Journal of Big Data*, vol. 8, no. 1, p. 161, 2021, doi: 10.1186/s40537-021-00548-z.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[11] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.

[14] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[15] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, 2020.

[16] M. Karmoude *et al.*, "Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks," *Science of The Total Environment*, p. 180593, 2025, doi: 10.1016/j.scitotenv.2025.180593.

[17] N. Fatima *et al.*, "Time series AQI forecasting using Kalman-integrated Bi-GRU and Chi-square divergence optimization," *Scientific Reports*, vol. 15, p. 12422, 2025, doi: 10.1038/s41598-025-12422-8.

[18] T. Tasioulis *et al.*, "The quest for the best explanation: Comparing models and XAI methods in air quality modeling tasks," *Applied Sciences*, vol. 15, no. 13, p. 7390, 2025, doi: 10.3390/app15137390.

[19] X. Zhou *et al.*, "Urban air quality management: PM2.5 hourly forecasting with POA–VMD and LSTM," *Processes*, vol. 13, no. 8, p. 2482, 2025, doi: 10.3390/pr13082482.

[20] B. Peng *et al.*, "Graph retrieval-augmented generation: A survey," *ACM Computing Surveys*, 2024, doi: 10.1145/3777378.

[21] Y. Ma *et al.*, "A survey of retrieval-augmented generation for large language models," in *IEEE International Conference on Information Technologies and Applications*, 2025, doi: 10.1109/ICTBAI68361.2025.00008.

[22] M. Bouni *et al.*, "Integrated IoT approaches for crop recommendation and yield prediction using machine learning," *IoT*, vol. 5, no. 4, p. 28, 2024, doi: 10.3390/iot5040028.

[23] D. Rajesh Kumar *et al.*, "On-device AI for climate-resilient farming with intelligent crop yield prediction using lightweight models on smart agricultural devices," *Scientific Reports*, vol. 15, p. 16014, 2025, doi: 10.1038/s41598-025-16014-4.

[24] F. M. Talaat *et al.*, "CardioRiskNet: A hybrid AI-based model for explainable risk prediction and prognosis in cardiovascular disease," *Bioengineering*, vol. 11, no. 8, p. 822, 2024, doi: 10.3390/bioengineering11080822.

[25] V. Gkouliaveras *et al.*, "Effects of climate change on health and health systems: A systematic review of preparedness, resilience, and challenges," *International Journal of Environmental Research and Public Health*, vol. 22, no. 2, p. 232, 2025, doi: 10.3390/ijerph22020232.

[26] C. De Guzman *et al.*, "Recent trends in research on the health impacts of climate change in the Philippines (2018–2023): A scoping review," *Frontiers in Climate*, vol. 7, p. 1594793, 2025, doi: 10.3389/fclim.2025.1594793.

[27] A. Samara *et al.*, "Impact of climate change and environmental adversities on maternal and fetal health," *Frontiers in Global Women's Health*, vol. 6, p. 1483938, 2025, doi: 10.3389/fgwh.2025.1483938.

[28] L. Gonah *et al.*, "The confluence of global climate change crisis, human health and health inequalities: A scoping review," *Global Policy*, vol. 17, 2025, doi: 10.52872/001c.137872.

[29] M. Shoaib *et al.*, "Air pollution and climate change: A dual threat to ecosystems and human health," *Journal of Environmental Challenges*, vol. 1, no. 1, 2025, doi: 10.62019/7w0pmn06.

[30] S. Bansal *et al.*, "Machine learning algorithms for predicting air quality index: A case study in urban and industrial zones," *Pollution Engineering*, vol. 13, no. 2, p. 354, 2025, doi: 10.21533/pen.v13.i2.354.

[31] R. Thangarajan *et al.*, "An AI based approach for smart air quality monitoring," in *IEEE International Conference on Contemporary Computing and Communication Technologies*, 2024, doi: 10.1109/ICCCNT61001.2024.10725962.

[32] A. Chinnaraju, "Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability," *World Journal of Advanced Engineering Technology and Sciences*, 2025.

[33] C. S. Reddy and M. Annamalai, "Explainable AI for diabetes risk assessment: Enhancing clinical decision support with interpretable machine learning models," in *IEEE International Conference on Contemporary Computing and Communication Technologies*, 2025, doi: 10.1109/ICSCSA66339.2025.11171123.

[34] I. A. Aderemi *et al.*, "Explainable AI for water quality monitoring: A systematic review of transparency, interpretability, and trust," in *IEEE Symposium on Reliability*, 2025, doi: 10.1109/SR.2025.3595500.

[35] F. El Guma *et al.*, "Early-warning prediction of visceral leishmaniasis mortality using a multivariate STL–deep learning hybrid approach on 20 years of monthly time series," *Frontiers in Public Health*, vol. 14, p. 1754966, 2026, doi: 10.3389/fpubh.2026.1754966.

[36] M. Hamzah *et al.*, "Emerging trends in statistical, machine learning, and deep learning models for air quality prediction: A bibliometric analysis," in *IEEE International Conference on Computing Technologies and Innovation*, 2025, doi: 10.1109/ICCTEIE66144.2025.11341833.

[37] S. Rafi *et al.*, "AI document search chatbot using retrieval-augmented generation (RAG)," in *IEEE International Conference on Communication, Power and Computing Technologies*, 2026, doi: 10.1109/I5CPS67958.2026.11452596.

[38] R. Yousefi Maragheh *et al.*, "ARAG: Agentic retrieval augmented generation for personalized recommendation," *arXiv preprint arXiv:2506.21931*, 2025, doi: 10.48550/arXiv.2506.21931.

[39] Z. Li *et al.*, "TutorLLM: Customizing learning recommendations with knowledge tracing and retrieval-augmented generation," *arXiv preprint arXiv:2502.15709*, 2025, doi: 10.48550/arXiv.2502.15709.

[40] Y. G. Nikam *et al.*, "AI and IoT integration in smart farming: Real-time monitoring and predictive analytics for sustainable agriculture," in *IEEE International Conference on Unmanned Aerial Systems*, 2025, doi: 10.1109/WorldSUAS66815.2025.11199045.

[41] Z. Han *et al.*, "An improved hybrid GC-LSTM framework for hourly nowcasting of ground-level NO2 concentrations over Beijing-Tianjin–Hebei region," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2025, doi: 10.1109/TGRS.2024.3514158.

[42] Y. Chudasama *et al.*, "Toward interpretable hybrid AI: Integrating knowledge graphs and symbolic reasoning in medicine," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3529133.

[43] P. Roy *et al.*, "Interpretable artificial intelligence for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge," *SAGE Open Nursing*, vol. 11, 2025, doi: 10.1177/20552076251327945.

[44] A. Alshraideh *et al.*, "Hybrid AI framework for the early detection of heart failure: Integrating traditional machine learning and generative language models with clinical data," *Cureus*, vol. 17, no. 6, 2025, doi: 10.7759/cureus.85638.

[45] P. K. Gupta *et al.*, "Data protection in enhanced explainable digital twins: Insights from hybrid AI," in *IEEE International Conference on Communication Technologies*, 2025, doi: 10.1109/CICTN64563.2025.10932525.

Behavior-Aware Access Control for IoT Networks Using Lightweight Machine Learning at the Gateway Level

Yaseen Alduwayl, Abdullah Alessa, Mounir Frikha
Department of Computer Networks and Communications,
College of Computer Sciences and Information Technology, King Faisal University, Al-Ahsa, 31982 Saudi Arabia

***Abstract*—The growing amount of heterogeneous devices with scarce resources is compromising the security of the Internet of Things (IoT), as they are more likely to adapt to a fixed and identity-based access control. Conventional security systems tend to assume that once a device is authenticated, the network may be exposed to credential theft, firmware, and insider abuse. In this study a behavior-sensitive access control solution is presented, which integrates lightweight supervised Machine Learning (ML) on the IoT gateway to provide dynamic authorization. Unlike the traditional passive intrusion detection models, the proposed framework uses a Supervised Random Forest model to process real-time statistical feature summaries in terms of mean, standard deviation, and sparsity of the IoT telemetry data. The method converts the output of anomaly detection directly into access (full, restricted or blocked) levels. The system was implemented on a Flask-based gateway and tested with ToN-IoT benchmark dataset. The results of the experiments show an anomaly-class recall of 0.9986 (99.86%) with 91,169 correctly detected attack and 125 false negatives among the 91,294 attack instances, for a security-oriented Zero Trust profile. As an example, when rounded to two decimal places, this value is 1.00, but the unrounded value is reported so as not to suggest 100% detection. The enforcement layer focuses on reducing risk and removes or filters out requests that were determined to be malicious or unauthorized in the scenarios. The architecture is designed to provide low latency through feature extraction and inference on the edge, which provides data privacy because telemetry processing is locally done without relying on the cloud.**

***Keywords*—Behavioral-aware access control; edge intelligence; gateway-based security; IoT security; intrusion detection; lightweight machine learning; random forest**

I. Introduction

The rationale behind the need of Secure Access Control in IoT is as follows.

Internet of Things (IoT) devices have greatly contributed to an increase in the attack surface of contemporary networks because of their rapid proliferation in the fields of smart homes, health care, industrial automation, and intelligent transportation. IoT systems are usually heterogeneous and resource-restricted gadgets that ceaselessly interact with gateways and cloud computing, usually devoid of human oversight. These systems allow automating and making decisions in real-time, but they also cause severe security threats in terms of unauthorized access, compromised devices, and insider abuse [1], [2].

Access control is a very important aspect of IoT security as it defines what devices can access network resources and services. Nevertheless, most current IoT implementations use static access control mechanisms that are based on rules and believe that a device will be trusted once it has been authenticated. This perception is progressively false in real-life situations where attackers are able to utilize stolen credentials, imitate legitimate devices, or utilize authenticated nodes and execute malicious activity [3]. Consequently, identity-based access control can not be effectively implemented by pure mechanisms alone to defend against the threats that are constantly changing IoT networks.

Recent research points to the importance of adaptive and situation-aware security measures that would continuously evaluate the behaviors of the devices, instead of using only initial authentication [4]. Specifically, dynamic access control that is applied to observed behavior can go a long way in minimizing the consequences of insider attack, compromised nodes, or privileged abuse. This impulse has led to the investigations on the behavior-sensitive access control systems which have incorporated intelligence directly at the IoT gateways to facilitate timely and effective enforcement of security.

A. Identity-Based Authentication Weaknesses

Sharing secrets, certificates, or tokens Identity based authentication mechanisms are common in IoT systems because they are very simple and do not require much computation. After an authenticated device is usually given some preset permissions, which do not change during its lifetime or session. Although this paradigm is effective, it implicitly presupposes that authenticated machines act in a benign manner, which is not always so in practice [5].

There are a number of security breaches which have shown that attackers could compromise identity-based controls through insecure credential management, device cloning, firmware vulnerabilities, or insider access [6]. In the context of that, a rogue device might be allowed to execute a full range of its permitted identity, yet still manifest malicious intent, e.g., stealing data, denial-of-service, and unauthorized access to resources. The traditional authentication systems find the detection of such behaviors as absolutely impossible since they do not observe the behavior of a device after authentication.

Also, statical authorization policies are not flexible enough to react to dynamic changes in the behavior of a device or network state. When permissions are received, they are not often removed without some form of manual intervention, so responses may be delayed and the attack period is very long

[7]. These restrictions have underscored the need to go beyond identity-based security models to adaptive mechanisms that keep reassessing the notion of trust on the basis of real-time behavioral observations.

B. Machine Learning and IoT Security

ML and IoT Security ML has become an important instrument to improve the security of the IoT, especially intrusion detection and traffic classification. Although deep learning models (e.g. CNNs and RNNs) can achieve high detection rates, they can be high in both latency and computational requirements, so that they are infeasible in resource constrained edge gateways. On the other hand, unsupervised schemes tend to confuse between complicated attack patterns and non-malicious anomalies. In order to overcome these drawbacks, this study capitalizes on lightweight supervised learning through a Random Forest classifier. The system is trained on the ToN IoT data and thus learns to differentiate between normal and attack patterns with high fidelity based upon statistical summaries of data streams. This method has a unique edge over deep learning in that it has low inference times that are conducive to real-time gateway enforcement, where detection coverage is also significantly better than that of basic threshold-based or unsupervised methods [8], [9], [10].

C. Research Scope and Organization of the Study

In this study, the design and test of a behavior-sensitive access control architecture of IoT networks is considered in which the authorization decisions are dynamically calculated according to real-time behavior of the devices. In contrast to the traditional methods which decouple access control and intrusion detection, the suggested system closely integrates lightweight ML with the authorization enforcement in the level of the IoT gateway. The low-latency and computational-overhead goals and deployability in the real-world IoT environment are the objectives of this design.

This study is devoted to the design, implementation, and evaluation of the behavior-aware access control architecture of the IoT networks. The major contributions of this work are:

- Benchmark-Driven Evaluation: Compared to the past literature, which uses synthetic data, the system is trained and assessed using the ToN-IoT dataset, which has solid performance in the presence of realistic heterogeneous sensor telemetry and network traffic.
- Statistical Feature Extraction: The addition of 5-dimensional feature extraction module, which is small in size, summarizes complex telemetry, facilitating the sparse feature processing in the edge.
- Zero Trust Enforcement: Zero Trust Enforcement: A Supervised Random Forest model is used to highlight the importance of a high anomaly recall and directly link the detection result to access-control enforcement. The robot model had a recall of 0.9986 when tested on the provided set of anomalies and the gateway only allowed restricted or blocked access when it detected high-level risk or unauthorized behavior.

The rest of this study will be structured in the following manner. Section II provides background on the IoT architectures, authentication models and gateway-centric security models. Section III will specify the problem, research goals, threat model, and contributions. Section IV explains how the selection of literature was performed and what shortcomings were found in the existing research. In Section V, the proposed system architecture and methodology are described. Section VI describes the experimental procedure, data and the results of the evaluation. Section VII and VIII address workflow issues, proposed workflows, and a comparative analysis. In Section IX, security is more widely discussed and more practically implications are explained, and the future research directions are discussed in Section X. Lastly, Section XI is the conclusion of the study.

II. Background and Context

A. IoT Architecture and Security Problems

The architecture of the Internet of Things (IoT) is most commonly divided into three basic layers, namely, the perception layer (sensing devices), the network layer (communication and data transfer), and the application layer (data processing and service delivery). In the real-world implementation, a second layer of gateway is used as an intermediary between the limited IoT devices and external networks or cloud services. The functions of this gateway include protocol translation, aggregation of data, authentication, and policy implementation [11].

As shown in Fig. 1 the layers of IoT architecture and outstanding the gateway between each layer.

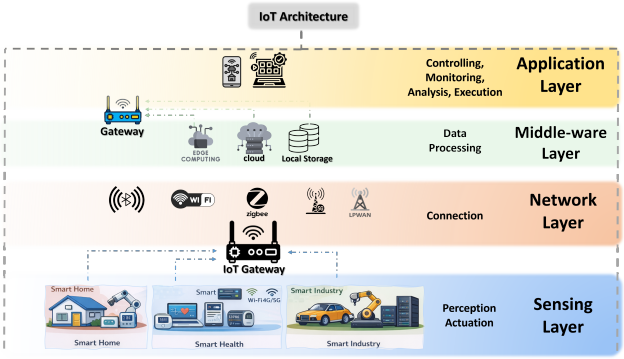


Fig. 1. IoT architecture layers (Sensing, Network, Middleware, and Application).

Although IoT has great architectural advantages, it presents serious security issues because it is heterogeneous in devices, has fewer computing capabilities, and it has many connections. Numerous IoT devices do not have the memory and processing power required to implement robust cryptographic algorithms or sustained security supervision and thus are targets when it comes to attacks [2]. Moreover, IoT networks are frequently installed in unmanaged or malevolent spaces and this raises the risk of physical interference, devices intrusion, and unauthorized access.

The other significant problem is the dynamic nature of IoT devices. The patterns of communication between the device

can vary with time and based on the circumstances of its operations, the frequency of requests and the services accessed. Conventional security systems cannot handle these types of behavioral differences, resulting in too much false warnings or unnoticed malicious breaches [3]. These issues indicate a necessity of a responsive, behavior-sensitive security system that functions near the network border [12].

B. IoT Authentication and Authorization Models

IoT security has some basic elements in authentication and authorization. Authentication is used to check the identity of a device and authorization is used to check what a device, which has been authenticated can do. In the IoT, the most commonly used authentication tools are pre-shared keys, digital certificates, and lightweight token-based authentication [5]. Such methods are also normally optimized around low overhead but exclusively identity verification.

In IoT settings, the process of authorization is typically realized with the help of such a static access control model as Role-Based Access Control (RBAC) or Attribute-Based Access Control (ABAC). Although these models are flexible in terms of the definition of permissions, they are generally based on predefined policies that do not change once authenticated [7]. Consequently, after a device has been authenticated and authorized then it is implicitly trusted until the end of its access session.

This assumption of static trust opens a big gap of security. In case a device is hacked after authentication, the traditional methods of authorization can not identify and react to the malicious activity in real-time. It has been observed that insider attacks and misuse of credential are some of the hardest threats to be detected by identity-based mechanisms alone [4]. Thus, continuous authorization models with dynamically re-evaluated trust according to dynamic behavior and not based on an identity are gaining interest.

C. IoT Network Intrusion Detection Systems

The topic of the Intrusion Detection Systems (IDS) has gained widespread attention as the method that helps strengthen the level of IoT security by detecting any malicious activities and detecting any unusual patterns of traffic. IoT IDS solutions can be broadly classified as signature-based, anomaly-based and hybrid solutions. These are anomaly-based IDS based on ML, which have become widely known thanks to their capability to identify unknown and zero-day attacks [8].

Recent studies on the application of deep learning models like convolutional neural network (CNNs), recurrent neural network (RNNs), and autoencoders to IoT intrusion detection have been examined. Although such models can be highly accurate in detecting objects, they can be expensive to implement in the IoT gateway because they often require heavy computational resources and GPU acceleration, making them unsuitable for deployment on resource-constrained IoT gateways [9]. In addition, a cloud-based IDS architecture creates extra latency and privacy issues since information is continuously being sent.

The other weakness of majority of IDS solutions is that they do not integrate with access control systems. Authorizations are normally made in real time without paying direct attention to detected anomalies which are usually logged or reported. Such a division between detection and enforcement slows down mitigation efforts and makes IDS as a whole less effective at attack prevention [10]. These drawbacks drive the adoption of lightweight anomaly detection being directly embedded into access control processes at the gateway level.

D. Gateway-Based Security Architectures

A new architecture of security Gateway based security has been developed as a viable way of implementing security policies in IoT networks. Centralization of security capabilities at the gateway allows devices with limited resources to avoid any complicated computations but receive extra protection [1]. On behalf of networked devices, gateways are able to monitor traffic, impose authentication, apply access control policy, as well as carry out anomaly detection.

Recent works support the concept of edge intelligence, in which ML models are executed on the gateway to examine device behavior on the device side. This technology lowers the latency and enhances scaling and maintains privacy of data over cloud-centric solutions [13]. Small ML models, including tree-based classifiers, unsupervised anomaly detection algorithms, etc., are especially applicable in this scenario because of their low computational footprint [14].

In spite of these, most of the current gateway-based security solutions continue to view intrusion detection and access plate as two distinct elements. Anomalies can be identified but authorization is not dynamically reconfigured according to behaviour identified. This vulnerability restricts the capacity of the gateways to be responsive towards the threats. Thus, the exploration of behavior-aware ML as a part of the authorization process at the gateway is an important milestone to more resilient and flexible IoT security systems.

III. Problem Statement, Objectives and Contributions

A. Problem Definition

Although authentication and access control mechanisms have been widely adopted in the IoT networks, the available solutions are still mostly stagnant and identity-based. After a device authenticates successfully with such credentials as pre-shared keys or tokens, it is normally issued with fixed permissions, which are not affected by any future actions. The given security model does not consider the possible cases of authenticated devices turning malicious because of credential theft, firmware attack, or insider abuse [5], [6], [15].

Fig. 2 illustate the procedure of risk that occured on IoT environment. Recent studies in the field of IoT security have paid much attention to the development of intrusion detection based on ML methods. Although such systems may be used to detect abnormal traffic patterns or malicious activities, they usually act as passive surveillance devices which do not have the option of enforcing these activities. Anomalies are captured and logged or reported, but do not happen instantaneously to influence authorization measures, leading to slow or inefficient mitigation [10], [16].

Moreover, a vast majority of AI-based IoT security algorithms are based on computationally intensive deep learning

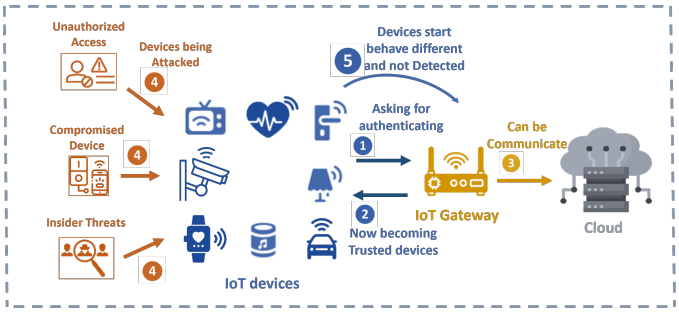


Fig. 2. Security challenges and behavioral anomalies in IoT environments.

architecture implemented on clouds. These methods have high latency, high bandwidth, and privacy risks, which complicate their practical use in the constraints of the IoT gateway resources [9]. Consequently, practitioner-oriented IoT deployment is facing a critical variation between behavioral threat identification and enforcement of dynamic access control [17].

B. Research Objectives

The main goal of this study is to develop and test a behavior-aware access control architecture of IoT network, which can dynamically modify the authorization decisions influenced by the online device behavior. However, in contrast to the traditional identity-related frameworks, the presented system keeps track of post-authentication actions and denies or blocks access in case of anchoring suspicious activity.

In order to accomplish this objective, the study aims at the following specific objectives:

- To examine the drawbacks of the in situ authentication and authorization systems in IoT-based systems and illustrate the fact that they fail to counteract insider and post-authentication attacks.
- To develop a lightweight ML-based behavioral analysis gateway-centric security architecture, not based upon cloud-based processing.
- To derive effective behavioural characteristics in the interaction of IoT devices that may be used to discriminate between normal and malicious activity successfully.
- To deploy a lightweight anomaly detection model to be used in real time at the level of an IoT gateway.
- To closely bind the results of anomaly detection to access control policies, so as to allow a response to be taken as soon as possible, that is, restricting or revoking access.
- To assess the proposed framework regarding the detection accuracy, the response time and the computational overhead.

All these are objectives to help close the divide between detection and enforcement but still be practical to the real-world IoT systems.

C. Assumptions and Threat Model

The study has its specific assumptions and realistic threat framework as per real-life IoT applications. The assumption is that the initial IoT device authentication with authentic credentials is never unsuccessful, and the gateway is not compromised. The communication between the devices and the gateway will be assumed to be secured with the help of the standard secure channels, including TLS [18].

The threat model takes into account the attackers who can interfere with the IoT devices once they have been authenticated. These enemies can use stolen credentials, duplicate legitimate nodes, or even inject malicious code on authorized nodes. The attacker can target to cause excessive traffic, misuse authorized services, gain unauthorized access to endpoints, or destabilize the normal processes of the gateway [3], [19].

The behavior that can be referred to as denial-of-service, abrupt fluctuations in the frequency of requests, atypical access patterns, and departures of learned behavioral patterns are all viewed as malicious activity indicators. Unlike rigid signature-based systems, the proposed framework utilizes a Supervised Random Forest model trained on the ToN-IoT dataset. By analyzing statistical feature summaries (such as mean, standard deviation, and range) rather than specific attack payloads, the system can generalize from the training data to detect malicious behaviors that statistically resemble known attack profiles, even if the specific execution varies.

This study is out of scope of physical attacks on the gateway, complex adversarial ML attacks, and large scale distributed botnets. The emphasis is still on the efficient behavior-aware authorization within the computational limits of one IoT gateway as shown in Fig. 3.

D. Summary of Contributions

This study brings the following main contributions to the study of the IoT security:

- Behavior-based Authorization Framework. It presents a new approach to access control whereby the authorization decision is continuously updated depending on how the device has acted as opposed to just based on identity alone.
- Lightweight ML on the Gateway. The system combines supervised, lightweight ML model, which can be applied to the IoT gateway, without depending on deep learning-based solutions on the cloud.
- Close Interconnection of Detection and Enforcement. In comparison to the traditional IDS methods, identified anomalies directly cause access control measures, allowing to restrict or revoke device privileges immediately.
- Practical and Low-Latency Design. The suggested scheme will be capable of working with low computational load and, therefore, suitable within a real-time IoT setting.
- Experimental Validation. An emulated IoT network is executed to test the framework using the ToN-IoT benchmark dataset to ensure fidelity. This allows the

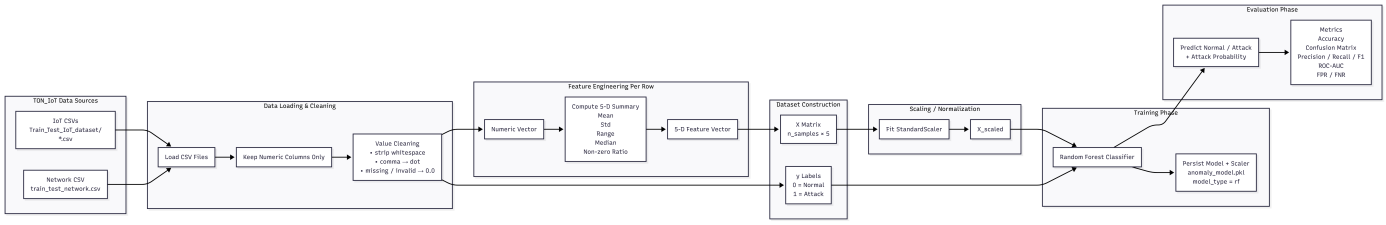


Fig. 3. Overview of the proposed Random Forest classification pipeline on the TON-IoT dataset.

system to be evaluated against realistic, heterogeneous sensor telemetry and actual network traffic patterns, covering diverse attack cases and performance parameters.

All these contributions cover the most important shortcomings in current IoT security offerings and show a realistic way forward to behavioral and adaptable access control.

IV. Literature Review Methodology

A. Literature Search Strategy

A literature search was carried out, which was systematic, to determine the current and topical studies in the field of IoT security, access control, intrusion detection, and ML-based anomaly detection. The main sources of literature searched were peer-reviewed journal articles and high-quality conference proceedings from 2020 or later to keep the review up to date with the latest research in IoT security. The earlier studies have only been added if they have been used as a basis for IoT architecture, historical developments in intrusion detection, or gateway-level security concepts.

Earlier publications like [11] and [20] and [21] are therefore not included as part of the current evidence base, but are kept as background or reference material. This distinction will not exclude seminal works, but will keep the focus on recent literature of IoT security.

A total of major academic databases were utilized such as IEEE Xplore, Elsevier ScienceDirect, SpringerLink, ACM Digital Library, and MDPI. The search queries were built based on the combination of terms, including the following: IoT security, access control, behavior-based authorization, intrusion detection, ML, anomaly detection, gateway security and edge intelligence. Refining took place with the help of the use of the Boolean operators to exclude irrelevant studies [22], [20].

Title and abstract screening were used to filter the first search results in order to make them relevant. A shortlisting was followed by full-text reviews on shortlisted papers to determine its congruency to the research interest of behavior-aware access control and lightweight ML in IoT settings.

B. Inclusion and Exclusion Criteria

There were clear inclusion and exclusion criteria that ensured that the literature review process was consistent and of good quality. The papers were to be included in case they met the following criteria:

- Published in 2020 or later, or published earlier but retained as a foundational work directly relevant to IoT architecture, intrusion-detection development, or gateway-level security.
- Specialized in IoT security, admission control, intrusion detection, or ML-based behavioral analysis.
- Security measures to be provided at the IoT, edge or gateway levels suggested or considered.
- Offered experimental assessment based on data sets, simulations or testbeds.

Articles were filtered out in case they fell into these categories:

- Published before 2020 and not required as a foundational or contextual reference for the architecture, threat model, or comparative baseline.
- Only concerned with cloud-based security and not with edge or gateway limitation.
- Only cryptographic authentication, without behavior analysis after authentication.
- Absence of experimental assessment or based entirely on theoretical discourse.
- Specific non-IoT networks like conventional enterprise or data center networks.

The filtering process ensured that the studies chosen were recent as well as relevant to the problem under investigation in this research.

1) Classification of selected studies: The chosen papers were divided into four broad groups in relation to the focus of research and the methodology applied:

- Models of Identity-Based Authentication and Authorization. These research works pay attention to the conventional authentication procedures like certificates, tokens, or attribute-based access control. Although they are useful in identity verification, they typically presuppose immobile trust following authentication and lack a concern about post-authentication conduct [7].
- Intrusion Detection Systems based on ML. The works in this category involve the application of supervised or unsupervised ML to identify malicious traffic or anomalous activity in an IoT network. Most of the solutions highlight the accuracy of detection but lack enforcement mechanisms [8], [9].

- Security Architecture based on Edge and Gateway. These researches suggest the use of security intelligence at the edge layer or IoT gateway to minimize latency and computation load. Nevertheless, the vast majority of practices consider intrusion detection and access control as independent entities [13].
- Behavior-Aware and Context-Aware Security Models. A more limited group of papers examines behavioral-based trust or adaptive security measures. Although these methods advance towards persistent authorization, they are usually based on complicated models or not include practical implementation specifications that can fit on lightweight gateways [4] [23].

This classification offers a systematic overview of available literature and the manner in which various methods deal with partial security concerns of IoT.

C. Limitations Revealed in Previous Studies

The discussion of chosen studies shows that there are a number of limitations, which are common and trigger the proposed research. To begin with, majority of identity based access control systems do not identify malicious activities that occur following authentication, which makes the IoT networks susceptible to insider and credential related attacks. Second, most ML-based IDS systems are detection-only and lack access controls which leads to slow or manual mitigation.

Moreover, much of the current literature is utilizing deep learning models that are implemented in the cloud, which adds latency, privacy implications and very high computational costs that cannot be implemented on the gateway. Even the gateway-centric solutions do not take into consideration the tight coupling between behavioral analysis and authorization enforcement.

Lastly, most of the proposed systems are evaluated solely through offline analysis of static datasets, decoupling detection from actual enforcement. This restricts their applicability to real-time IoT systems, where authorization decisions must be made instantaneously. The combination of these limitations shows that it is necessary to have a lightweight and behavior-aware access control system that is integrated directly into the IoT gateway to combine detection with immediate authorization decisions.

D. Related Work

The most recent studies on IoT network protection have discussed ML, edge computing, and access control models. Although such studies have enhanced the level of detection and efficiency of the system, there is still a fundamental gap of combining real-time behavioral monitoring with adaptive access control per gateway. This field has 20 recent contributions, the discussion of which is critically compared below.

E. Machine Learning-Based Intrusion Detection in IoT

In [24] the authors have reviewed and proposed some of the ML techniques to optimize the IoT in regard to security, authentication, access control, and malware detection. They also discussed spoofing, jamming and eavesdropping and the goal was to prove that ML may be utilized instead of conventional security in dynamic settings. The study is about generic IoT applications and the examples of smart grids and vehicular networks. Regarding methodology, they used game-theoretic modeling and simulation to evaluate different strategies in ML such as Q-learning and supervised learning. They used performance measures as detection accuracy, false alarm, utility of a system and demonstrated that Q-learning could reduce authentication error by approximately 10-15 percent compared to traditional schemes in their simulation. The key findings are that, though ML can be utilized to improve security, it introduces overhead, which can be challenging to handle using low-power devices. This is not a fully appropriate solution to your provided problem as, though it deals with ML to provide access control, it relies on reinforcement learning (Q-learning) and game theory models, which are computationally expensive and can require long convergence times, which is not ideal in resource-bounded gateway processing.

In this study [12], the authors suggested a framework to identify cyberattacks of smart city apps with the help of different ML algorithms. They have tackled the problem of security of smart cities, i.e., detecting anomalies of sensor and communication systems. It is an intelligent urban application. Their approach was to train and test classifiers such as Decision Trees, Random Forest, and k-NN on a dataset they created. They used accuracy, precision, recall, and F1-score and noted that Random Forest achieved 99.9% accuracy with 0.1% false positive. The findings show that Random Forest is powerful with smart city traffic. They used a dataset created by a simulated smart city testbed with 362,377 samples. This solution uses Random Forest but is not suitable to your problem because it is not an active IDS and does not tie detections to real-time enforcement actions at the gateway level.

The author in the article of [9] proposed a Deep Learning-based hybrid Intrusion Detection System (IDS). Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) were employed by them to identify sophisticated attacks. They worked on improving the ability to detect attacks like DDoS and ransomware. Particle Swarm Optimization (PSO) was used to optimize the model. They evaluated their model with accuracy, precision, recall and F1-score with 99.64 Bot-IoT and 99.45 TON_IoT dataset. The hybrid CNN-LSTM model was more effective compared to the single models. This version is inappropriate as it relies on a computationally-intensive Deep Learning architecture (CNN + LSTM) which uses processing power and cannot be implemented on an IoT gateway with limited resources. Furthermore, it is a detection system (IDS) only, which lacks an access control mechanism that provides an automatic implementation of blocking policies.

In [17], the authors introduced an ensemble Intrusion Detection System (IDS) that uses flow statistical features to detect malicious traffic. They concentrated on identifying sophisticated attacks in the IoT networks by combining classifiers (Naive Bayes, Decision Tree, and Artificial Neural Network). They produced statistical features and employed a voting - based ensemble classifier. They measured the performance with Accuracy, False Alarm Rate, and Detection Rate reaching 99.5% detection rate with low false alarms. The findings indicated that ensemble method minimized errors when compared

to single classifiers. This is also inappropriate as it employs an ensemble, which is more computationally expensive than one Random Forest model and is still a passive IDS, which does not enforce anything in real-time.

In [10], the authors applied an ensemble ML approach to detect DDoS attacks in the IoT network. They trained algorithms (Random Forest, Extra Trees, XGBoost) and voted on them with a soft voting classifier. They compared the performance based on Accuracy, Precision, Recall, and F1-score with a 99.7 accuracy and 0.99 F1-score. The findings revealed that ensemble approaches are more consistent than individual models. This proposal is not applicable because it concentrates only on DDoS and has no behavior-aware mechanism of access control to be used in real-time.

In the study by [3], the authors performed the review of Intrusion Detection Systems (IDS) in IoT with the focus on Multi-Access Edge Computing (MEC) and ML. They stressed the idea of moving security functions to the edge in order to minimize latency and bandwidth. They used a method of a systematized literature review that categorized IDS based on the deployment and detection method. Some of the performance measures they reported included performance in terms of detection and processing time. The results indicate that MEC-based IDS can decrease the response time but might not be scaled to heterogeneous IoT data. The article is inappropriate as it is a review of architectures and does not give an actual system or a behavior-conscious access control implementation.

The authors of the article in [15] proposed an Intrusion Detection System (IDS) using a supervised ML to detect malicious behavior in IoT networks. They trained classifiers (Random Forest, Decision Tree, k-NN and Naive Bayes) with the UNSW-NB15 dataset. They assessed the performance by Accuracy, Precision, Recall, and F1-score with the accuracy of the performance in the case of Random Forest and Decision Tree being 98.9% and 99.5%. The findings revealed that tree-based models are correct and less expensive to compute. This is not a solution to you because it is a passive IDS which lacks automated access control measures at the gateway.

The authors of the article in [4] evaluated Intrusion Detection Systems (IDS) based on the usage of ML in Ioot. They compared the classifiers: Random Forest, SVM and Ensemble methods between the datasets. They employed Detection Rate, False Positive Rate and Accuracy with ensemble techniques showing detection rates greater than 99%. The findings indicated that the performance is dependent on the dataset. They were used in NSL-KDD, CIDDS-001 and UNSW-NB15. This solution is not applicable to your problem as it is benchmarking research and lacks the integration of detection and real-time authorisation.

In [25] the authors enhanced the Isolation Forest algorithm by incorporating it with the k-means clustering algorithm to identify anomalies. They followed a two-step process, in which k-means splits data into clusters and Isolation Forest identifies outliers. They tested the performance in terms of AUC, accuracy and processing time with better AUC than the traditional iForest. The findings indicated a better detection of local anomalies at the expense of increase in computational cost. They were synthetic datasets and KDD Cup 99. This is inappropriate because it is an unsupervised anomaly detector and it does not offer dynamic access control.

F. Deep Learning and Network Traffic Analysis in IoT

The authors in [26] developed a Deep Learning model to identify botnet attacks on healthcare IoT networks. A Deep Belief Network (DBN) that was trained on the Bot-IoT dataset was used. They assessed the performance based on accuracy, precision, recall, and F1-score, obtaining an accuracy of 99.25% and a recall of 98.8%. These findings indicated that Deep Learning has an application in monitoring botnet activity. This does not fit well as it involves Deep Belief Network, which is computationally intensive and cannot fit a resource-constrained IoT gateway.

The authors of the article in [7] studied the detection techniques of the IoT botnets. They contrasted signature-based, anomaly-based and DNS-based techniques. It was a comparison literature review procedure. They used detection rates greater than 99 percent and false positive rates to evaluate methods. The results indicated that ML methods are prospective but they are very expensive to process. They mentioned datasets like N-BaIoT. This is not a solution to your problem because it is a survey and lacks an integrated mechanism of access control to automatically block at the gateway.

These authors in [6] provided a case study of the Mirai botnet that led to DDoS attacks. They evaluated the Mirai scan, infect and organization of IoT devices to attack with more than 1 Tbps. They analyzed the source code and pattern of attacks of Mirai. They announced the size of attacks (e.g., 620 Gbps) and compromised devices (more than 600,000). The results revealed the use of weak default credentials. This solution is inappropriate as it is an analysis of an attack and not a defense ML or gateway-based solution.

G. Surveys and Emerging Technologies in IoT Security

The authors in [27] surveyed the problem of security and forensics in the IoT ecosystem. They categorized threats in each of the layers of IoT and addressed the problem of data privacy, authentication, and standardization. They looked at the available literature and offered security requirements in form of a taxonomy. The results revealed the gaps in the IoT forensics and difficulties associated with the resources limitations of devices. The study is inappropriate as it is a theoretical survey and lacks practical access control solution based on ML or gateway.

The researchers in [28] have carried out a survey to categorize security challenges and solutions in the IoT. They discussed the vulnerabilities, data privacy, and authentication in various areas of IoT. They were able to review over 100 articles which lacked experimental analysis or performance measurements. These results highlighted the importance of light-weight security as a result of limited resources. The study is not appropriate as it is theoretical and lacks a real-world architecture and behavior-based access control mechanism.

In this study [29], the authors found out the use of Federated Learning (FL) along with IoT to address centralized ML constraints. They solved the problem of data privacy, latency and bandwidth because they allowed distributed training on

edge devices. The areas included in the study were IIoT, IoV, healthcare, and smart cities. The research design was literature review that categorized the approaches into resource management, privacy, and security. The results indicated that FL enhances privacy but lacks in data heterogeneity and device limitations. This solution cannot be used as it has to be decentralized in terms of coordination and is not compatible with a centralized lightweight gateway-based solution.

The authors of the study in [30] reviewed the application of Federated Learning (FL) to enhance cybersecurity in IoT. They tackled data privacy and centralized processing through collaborative detection without transfer of raw data. The research design was a theoretical analysis of FL architectures. They talked of accuracy, communication overhead, and convergence time, with comparable accuracy to centralized models with less bandwidth. The results indicated weaknesses like susceptibility to poisoning attacks and expensive communication. Your problem is not well addressed using this solution as the solution is based on a distributed training process as opposed to a centralized lightweight gateway-based solution.

In the study by Li and et al. [13] they made a survey of Edge Computing within the Internet of Everything (IoE) was conducted. They also resolved the problem of latency, bandwidth, and privacy of cloud-based processing, demonstrating that edge computing decreases latency. The research design was a literature review with no experimentation. The results indicated that edge computing lowers the latency but presents a set of problems including resource limitations and security control. This is not helpful to your problem as it is a theory overview and does not offer a practical access control algorithm or behavior-based security.

The researchers in [1] performed an IoT security, privacy and trust review. They talked about scalable middleware gaps, secure data management, and authentication gaps. A literature review that categorizes security challenges was used as the methodology. The results revealed the absence of lightweight and single security models. This does not apply as it is a descriptive survey and does not give a practical implementation to be applied to real-time blocking at the gateway.

H. Access Control and Authentication Mechanisms in IoT

Within the framework of this general survey [31], The researchers analyzed the use of blockchain with IoT to enhance access control. Among the problems that they solved were centralized control and single point of failure, which sought to have a decentralized structure. The research design was a survey that reviewed and classified blockchain-based solutions. The results of the study was revealed that blockchain enhances security but bring with latency and computational overhead. Your problem does not fit this solution, as it is both a heavy and decentralized scheme with delays, and is not suitable to a lightweight real-time gateway design.

The researchers in [32] delivered a lightweight authentication system to smart home devices with resource-constrained capabilities. They dealt with big computational cost and attacks like replay attacks and man-in-the-middle to facilitate secure communication at low energy consumption. The security analysis and performance comparison methods were used. They compared the cost of computation, the cost of communication and the energy consumed, which demonstrated a lower cost at the expense of security. It is not a valid solution as it remains an authentication protocol and does not deal with post-authentication behavior or compromise devices identification.

The researchers in [33] have developed an access control model of Electronic Healthcare Systems based on trust. They dealt with the dynamic control of access by calculating the trust score using user behavior. The technique involved user behaviors to modify trust and access privileges. They tested the success rate of access, time to compute trust and penalties to behavior. The findings demonstrated that the model is able to adjust to changes in behavior. This does not fit as it is human user-based and complicated computation of trust, rather than device behavior at a real-time IoT gateway.

I. Lightweight and Edge-Based Security Solutions for IoT

As the researchers state [34], the multimedia traffic of IoT is suggested to be provided by the so-called media-aware security architecture. To solve this trade-off between security and QoS they adjusted security according to delay and jitter. The testbed methodology involved a delay, jitter, packet loss and throughput simulation testbed. The findings demonstrated better adaptive security QoS. This does not apply as it prioritizes multimedia traffic and sacrifices security in favor of QoS rather than providing a strict security enforcement of the general IoT devices.

The researchers in [21] suggested an inexpensive flow-based security architecture of smart home IoT devices. They modeled the behavior of devices based on network traffic without deep packet inspection. The method involved actual traffic of the IoT devices to create typical behavior patterns. They tested the accuracy of classification and the accuracy of anomaly detection. The findings indicated that IoT devices are easy to profile and their traffic is predictable. This is not a solution as it is a passive monitoring tool and does not give automated access control to prevent malicious devices (Table I).

J. Summary of Gaps Identified

The current literature is characterized by the following limitations:

- The majority of solutions revolve around detection and not prevention.
- ML-based IDS are not always coupled with access control.
- IoT gateways cannot support heavy deep learning models.
- Behavior-based authorization that is real-time is poorly explored.

These loopholes are the direct inspiring factors of the proposed behavior-aware access control framework that brings together the lightweight ML with real-time enforcement of authorization at the IoT gateway tier.

TABLE I. ANALYSIS OF RELATED STUDIES

Ref	Year	ML/DL Models	Dataset	Results	Key Findings	Limitations
[4]	2021	Decision Tree, Random Forest	Smart Home / IoT Traffic	High accuracy; offline analysis	Effective for smart home and city profiling	Offline analysis only; no real-time integration
[3]	2021	LSTM (Deep Learning)	IIoT Traffic	High detection accuracy	Suitable for large-scale IIoT	Cloud dependency; high latency
[10]	2021	Isolation Forest (Unsupervised)	IoT Traffic	Effective anomaly detection	Works well for unseen DDoS attacks	Independent IDS; cannot prevent attacks
[9]	2021	CNN + RNN (Hybrid DL)	Healthcare / Industrial IoT	High detection rates	Good performance in heterogeneous IoT	High complexity; no adaptive revocation
[13]	2022	Lightweight ML + Edge Intelligence	IoT Edge Traffic	Reduced response time	Scalable edge-based security	Detection and access control are separate
[23]	2022	Statistical Trust Scoring	Healthcare IoT	Improved authorization decisions	Trust-based decisions enhance insider resistance	Historical data only
[15]	2021	Clustering-based ML	IoT Traffic	High flexibility	Effective for unknown anomalies	High false positives; no ACL feedback
[30]	2021	Edge Traffic Filtering	IoT Gateway Traffic	Reduced backend load	Efficient local filtering	Signature-based; not adaptive behavior
[36]	2022	Distributed Edge Authorization	IoT Edge Networks	Lower latency	Effective low-latency access control	Static policies; no behavioral scoring
[40]	2020	Lightweight Authentication	IoT Devices	Preserved battery life	Efficient initial handshake	No post-authentication anomaly handling
[17]	2021	Ensemble ML Models	IoT Traffic	Increased detection accuracy	Reduced classification errors	Computationally heavy
[16]	2020	Distributed Deep Learning	IoT / Botnet Traffic	Improved botnet detection	Localized edge learning benefits detection	Complex models; high compute requirement
[1]	2020	Survey / Theoretical Analysis	N/A	Identified security challenges	Highlights trust and privacy gaps	No implementation/testing
[2]	2020	Deep Learning + Big Data	IoT Traffic	Improved detection metrics	Useful for large-scale IoT traffic	High resource usage; cloud reliance
[5]	2021	ML-based Botnet Detection	IoT / Mirai Traffic	Accurate botnet detection	Recommends future research directions	Reactive approach; no adaptive access control
[6]	2021	ML + MEC Framework	IoT Traffic	Efficient anomaly detection	Edge-level anomaly detection	Limited dynamic access control integration
[7]	2021	Botnet Pattern Analysis	IoT Networks	Detection of Mirai variants	Highlights IoT network threats	Detection only; no adaptive enforcement
[8]	2021	CNN + RNN Supervised Learning	IoT Traffic	Improved detection	Good for heterogeneous IoT	Centralized; heavy computation
[11]	2020	Cluster-based Isolation Forest	IoT Traffic	Efficient anomaly detection	Works without labeled data	No access control integration
[12]	2022	Lightweight ML at Edge	IoT Edge Traffic	Reduced latency; privacy preserved	Edge-level enforcement feasible	Detection and access control disconnected

V. METHODOLOGY

A. System Architecture Overview

The suggested system is based on the gateway-centric security architecture where all the interactions with the IoT devices are facilitated by a centralized gateway. The gateway can be used as the point of enforcement of authentication, behavior monitoring, anomaly detection, and access control. In this design, the computational load on IoT devices with limited resources is reduced, and real-time security decisions are made near the network edge.

It is composed of four closely coupled modules: 1) authentication, 2) behavioral monitoring, 3) lightweight ML-based anomaly detection and 4) access control enforcement. Devices are constantly checked after successful authentication and their behavior is assessed with the help of the lightweight anomaly detection model. ML model output has a direct impact on making decisions on authorization, which enables the system to enable and limit or deny access dynamically as shown in Fig. 4.

Designing these elements at the gateway can help the system to avoid the latency and privacy problems that come with cloud-based intrusion detection but at the same time, the detected anomalies are immediately converted to enforcement measures.

B. IoT Network and Gateway Model

The IoT network model presupposes the communication between a number of heterogeneous devices and a single gateway through the REST-based communication. Devices can initially register and identify to the gateway through a unique device identifier and a shared secret. When a device has been successfully authenticated, it is first assigned full access permissions and behavior monitoring is then initiated [28].

The gateway stores state data of each device, such as authentication status, access level, and history of recent interaction. The gateway becomes the only access control decision maker since all device requests are funnelled through it. This model of a centralized gateway is representative of a real-life IoT implementation that is usually used in smart home, industrial, and healthcare settings [1], [24].

The gateway also determines the various access levels, which include full, restricted, and blocked and are related to a predetermined list of authorized endpoints. This allows the capabilities of devices to be easily controlled to fine grains and also enables the gradual decrease of capabilities instead of abrupt disconnection to enhance system robustness and usability.

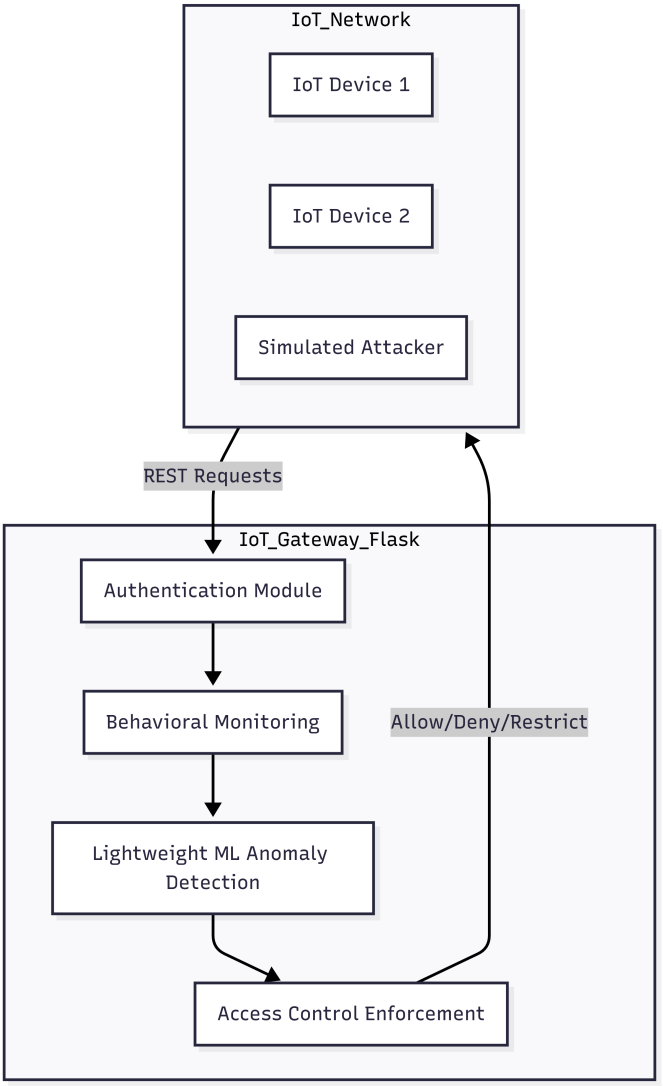


Fig. 4. Gateway-centric behavior-aware access control architecture.

C. Extracting Behavioral Features

In order to deal with the heterogeneity of IoT traffic, in which different sensors (e.g., thermostats, GPS trackers, Modbus devices) produce radically different data payloads, the system no longer inspects raw payloads, instead promoting statistical summarization. A fixed dimension 5-dimensional feature vector is obtained, such that each row or data window processed by the gateway is described by the current behavioral state of the device. [35], [36].

These features are: As can be seen in the analysis of the ToN-IoT dataset, they are:

- Mean: The arithmetic mean of the numeric values in the telemetry row which gives a baseline of the signal magnitude.
- Standard deviation (std): Indicates the degree of volatility or stability in data stream. Large deviations can be associated with unpredictable behavior as found in flooding attacks or malfunctioning sensors.
- Range: The maximum value less the minimum value and is used to measure the spread of the data.
- Median: The center of interest of the data used to reduce the skewing effect of the outliers.
- Non-zero Ratio: Fraction of the values in the row that are non-zero. This sparsity measure is important in detecting attacks that either flood the fields with zeros or, alternatively, flood the otherwise empty fields with data.

It is a 5-D representation scale of the input to the ML model, so that the computational cost of extracting the features is low irrespective of the type of device.

D. Lightweight Supervised Model Selection

The detection engine is a supervised random forest classifier at its core. It was chosen instead of unsupervised Isolation Forests because there are labeled data (Normal vs. Attack) in the ToN-IoT dataset which enables the system to learn particular limits of deviant behavior, as opposed to mere outliers. [25]. The pipeline of training is as follows:

- Data Loading: The system loads the ToN-IoT training set, a non-homogeneous set of datasets that consists of IoT and Network datasets.
- Preprocessing: The training data is then preprocessed using a StandardScaler to normalize the 5-D feature vectors, so that those with larger values (such as the one labelled Range) do not dominate the learning mechanism.
- Training: The Random Forest classifier is trained by the entire labeled data (28706 normal samples and attack samples). The model is trained to estimate the non-linear association among the 5 statistical features and the binary class labels (0: Normal, 1:Attack).
- Serialization: The trained model and scaler are pickled (serialized) to be deployed on the Flask gateway, which allows inference times of sub-milliseconds in live mode.

E. Data Loading and Preprocessing

The system relies on the ToN-IoT benchmark dataset that includes the files of IoT telemetry and network traffic in CSV. In order to deal with heterogeneity of these sources, there is a strict filtering procedure based on numbers only.

- Column Filtering: Non-numeric values, e.g., timestamps, IP addresses, and protocol strings, are filtered out so that the model is lightweight and geared to statistical behavior.
- Data Cleaning: Raw data is purged by removing the whitespace and standardizing the decimal separators.
- Null Handling: Null values or invalid placeholders are represented as 0.0 to have a continuous numeric vector per row.

F. Statistical Feature Engineering

Each cleaned telemetry row or data window is converted to a fixed 5 dimensional feature vector containing the mean, standard deviation, range, median and ratio- where a data row is not all zeros. See Section V-C for details. The goal of this subsection is to avoid duplication, so it only discusses the impact of the previously defined representation on the post-pre-processing. The cleaned numeric values are converted to a standard 5-D feature space, and no matter how many numeric features the data source of each IoT contains, the dimensions of the resulting matrix are all the same, which can be used for scaling and classification using Random Forest.

G. Model Training and Scaling

The pipeline of learning under supervision transforms the processed 5-D vectors to a labeled training dataset on Fig. 5 showing all steps of implementation including the 5-D vectors as part of feature engineering.

- Normalization: A `StandardScaler` object is applied to the feature matrix X, so that all five statistical features are in a similar range. This ensures that features with larger ranges do not dominate the learning process.
- Random Forest Training: A supervised Random Forest classifier is trained on 28,706 normal and attack samples to learn the non-linear boundaries between benign and malicious behavior.
- Model Serialization: The trained model and scaler are pickled (`anomaly_model.pkl`) to be deployed in real-time into the Flask-based gateway.

H. ML and Access Control Mechanism Integration

The major contribution of this study is that there is a close link between ML-based anomaly detection and access control enforcement. In contrast to the conventional IDS models that are put in place separately, the anomaly score generated by the ML model directly changes the level of authorization of the device [21].

There are two threshold values, which are used to categorize behavior as normal, suspicious, and malicious. Normal devices are not restricted, whereas suspicious devices get access but with limited permissions and rate limits. A gateway is accessed to block unscrupulous actions and revoke access immediately.

This integration will provide the security responses to be proactive and automated that reduces the time frame of attack and removes the aspect of manual intervention. The system directly incorporates intelligence into the authorization pipeline to provide behavior-sensitive access control which can be used in real-time IoT contexts.

VI. Experiment Preparation and Findings

A. Simulation Environment and Tools

The proposed behavior-aware access control framework was experimentally tested with the help of the simulated IoT setting in Python. The simulation was created in a way that it was modeled after real-life gateway-based IoT deployments, whereas the behavior of each device and attack situations could be fully controlled.

TABLE II. ToN-IoT Dataset Specifications: Rows, Raw Columns, and Feature Extraction

Dataset Component	Max Rows	Raw Cols	Numeric Cols	Final Features
IoT Fridge	15,000	6	2	5
IoT Garage Door	15,000	6	2	5
IoT GPS Tracker	15,000	6	2	5
IoT Modbus	15,000	8	4	5
IoT Motion Light	15,000	6	2	5
IoT Thermostat	15,000	6	2	5
IoT Weather	15,000	7	3	5
Network Traffic	15,000	44	∼23	5
Combined Dataset	**120,000**	**Varies**	**Varies**	**5 (Fixed)**
Normal Samples (Train)	28,706	–	–	5
Total Samples (Eval)	120,000	–	–	5

The IoT gateway was deployed on Flask framework in which the framework presented RESTful APIs to allow level of registration of devices, authentication, processing of requests and implementing access control. NumPy and scikit-learn libraries were used to provide lightweight ML and data processing. Supervised Random Forest algorithm and normal feature scaling methods were used in detecting anomalies. Other tools like matplotlib were employed in visualising the results and in analyzing performances.

The experiments were performed on a normal computing platform without any hardware accelerations, and the fact that it was possible to run the suggested system on devices with limited resources indicates the presence of IoT gateways that can be used to implement the system. This toolchain represents the software elements that are found in the real-world IoT edge settings [13].

B. Dataset Specifications

The framework utilizes the ToN-IoT benchmark dataset, which contains heterogeneous telemetry from diverse IoT sources and network traffic. To ensure a balanced and representative training environment, the system processes seven distinct IoT device logs and one comprehensive network traffic dataset.

As the IoT devices produce varied data payloads (e.g., temperature vs. Modbus registers), a strict numerical extraction policy is applied. Categorical fields such as IP addresses, protocol strings, and timestamps are excluded to focus the model on statistical behavior. Table II provides the detailed specifications of the rows and columns processed from each source.

1) Data processing and dimensionality reduction: Regardless of the original number of numeric columns (ranging from 2 in simple sensors to 23 in network traffic), every raw row is mapped to a standard five-dimensional (5-D) feature space. This ensures interoperability across the gateway. The mapping process involves cleaning missing values (replacing them with 0.0) and calculating the five summary statistics: mean, standard deviation, range, median, and non-zero ratio.

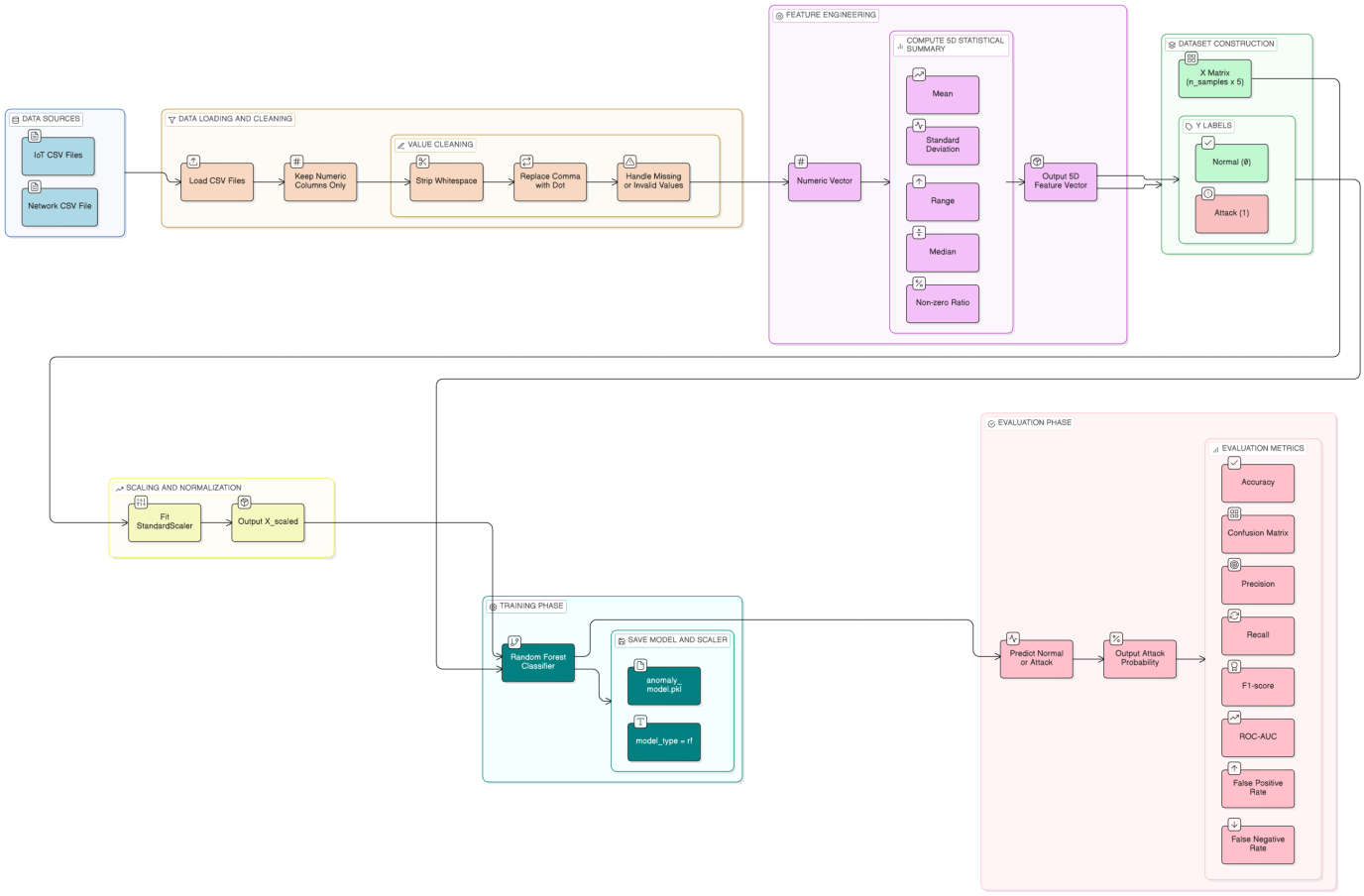


Fig. 5. End-to-end supervised random forest training pipeline: From raw ToN-IoT data loading and 5-D statistical feature engineering to model serialization and evaluation.

C. Dataset and Experimental Setup

The use of the ToN-IoT dataset as the ground truth to prove the performance of the system in a realistic threat scenario was applied. This dataset has been widely known in determining attacks in heterogeneous IoT networks [37].

- Training Phase: It was trained on a smaller set of samples where it involved 28,706 normal samples and attack samples where the Random Forest was able to develop a full spectrum of a decision boundary.
- Evaluation Phase: The system was assessed with regard to a separate assessment set of 120,000 total samples. Such a large-scale test was rich in diverse attack types (e.g., DoS, DDoS, scanning, and ransomware) and non-attack traffic, which was a stringent stress test of the classifier.

D. Performance Metrics

The proposed system was tested in terms of several performance measurements to measure the performance in detection and the feasibility in practice. To identify the general correctness of the anomaly classification, detection accuracy was measured. The trade-off between false positives and false negatives, especially with regard to detecting malicious behavior, was determined by computing precision, recall and F1-score.

Furthermore, receiver operating characteristic (ROC) analysis and area under the curve (ROC-AUC) to measure the capability of the model to differentiate between normal and abnormal behavior at various threshold values were also applied. The confusion matrix analysis was used in order to obtain the true positive, false positive, true negative and false negative rates.

To gauge the level of practical suitability, the response time and the computational overhead were also monitored qualitatively, with respect to the capability of the gateway to handle requests and make access decisions in real time. Such metrics can be considered an overall assessment of the effectiveness of security and efficiency of the systems [8].

E. Results and Analysis of the Experiment

Findings and Discussion The system performance was evaluated with the help of the standard classification metrics, and it was found that the system has a peculiar zero-trust security profile.

- Accuracy and Overall Performance: The model has a total Accuracy of 81.17% over the 120, 000 samples

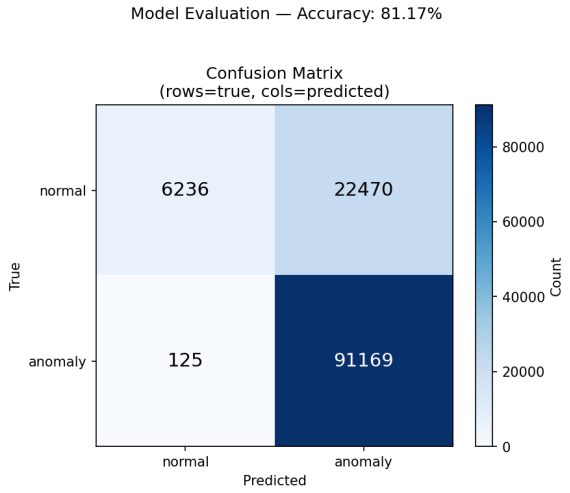
of evaluation. Although this is a competitive accuracy, the distribution of errors (False Positives vs. False Negatives) is the most important observation that is necessary in a security application.

- Recall: Zero-one anomaly-class recall of 99.86% was obtained by dividing the number of correctly detected attack samples by the total number of attack instances, which were 91,169 and 125, respectively. It is reported with two decimal places throughout the manuscript, but is rounded to 1.00 for the purpose of discussing the value of this determination. This result confirms that the framework is able to identify almost all the attack samples evaluated and that any false negative results should not be interpreted as a result of complete or perfect attack detection.
- Precision and False Positives: The precision in the anomaly classes was 0.80. The confusion matrix is reporting 22,470 false positives, which means that there were many more legitimate samples that were incorrectly identified as anomalous. This is a situation with a security conservative operating point, a situation in which the model is security-oriented, meaning that it minimizes false negatives while allowing for more false alarms. In real-world IoT implementations, these false positives can temporarily deny access to legitimate devices, delay telemetry reporting, burden the operator's workload, and decrease the trust in the access control system by the user. To alleviate this effect, anomalous cases with low confidence should not be blocked outright but instead should be placed in a limited-access mode; device-specific thresholds, periodic retraining, whitelisting of safety-critical telemetry, second-stage verification can be employed to minimize unnecessary disruptions.
- ROC-AUC: The model has a ROC-AUC score of 0.82, which represents a profound capacity to arrange randomly selected positive examples higher than negative one.

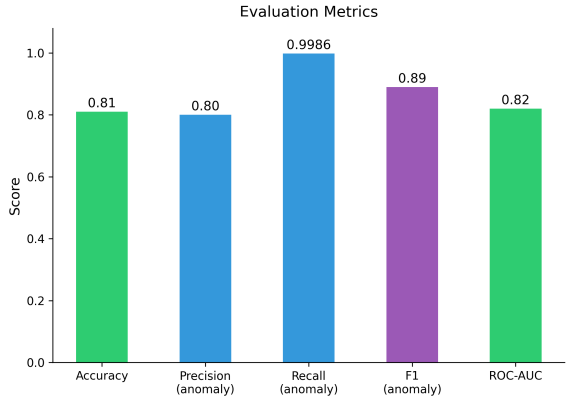
On Fig. 6 illustrates the confusion matrix and the main performance metrics, including accuracy, F1-score, and ROC-AUC.

Operationally, the enforcement of access controls was implemented as soon as an anomaly was detected to prevent the rapid restriction or revocation of device permissions. The model was lightweight and, therefore, provided low computational cost and insignificant effects on the response time of gateways. These findings confirm the feasibility of deploying lightweight ML at the gateway level with the access control and makes sure that the offered framework meets the security-efficiency balance [34].

The evaluation was conducted through a controlled execution script designed to stress the gateway across four distinct phases. First, baseline latency was established using 25 sequential requests from a single device with a 0.1s delay to capture the minimum, maximum, and average processing times reported by the gateway. Second, "Normal Operation" metrics were gathered by simulating 5 devices under a round-robin load for 10 seconds, alternating between standard requests and



(a) Confusion matrix of the random forest model.



(b) Performance metrics (Accuracy, F1-Score, ROC-AUC).

Fig. 6. Evaluation results of the random forest model.

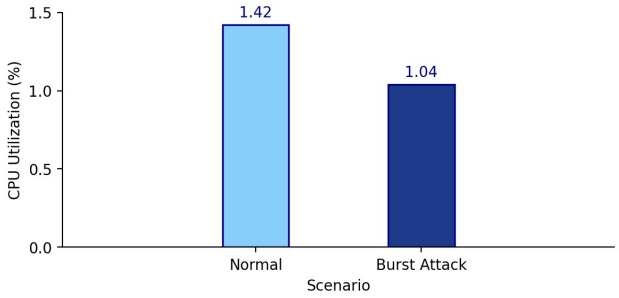


Fig. 7. CPU Utilization under normal vs. burst attack scenario.

administrative status checks. Third, a "Burst Attack" scenario was emulated by a single device issuing 150 requests at high frequency (0.02s intervals), with CPU and RAM samples recorded every 10 requests. Finally, scalability was tested by registering varying device counts (n=1,5,10,20,50) and calculating the average gateway-reported latency across 10 requests per device. These metrics were then combined into a JSON object, which was then processed to create the CPU-utilization, memory-usage and average-processing-latency visualizations

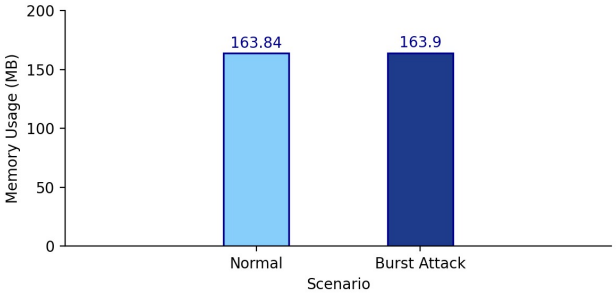


Fig. 8. Memory usage under normal vs. burst attack scenario.

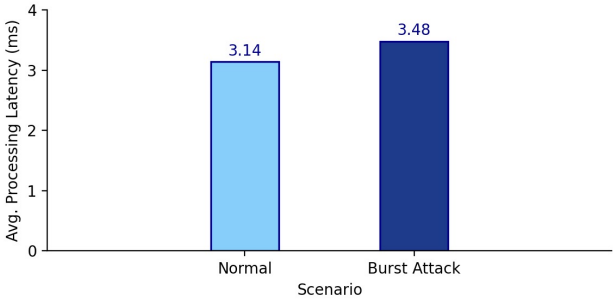


Fig. 9. Average processing latency under normal vs. burst attack scenario.

TABLE III. GATEWAY COMPUTATIONAL OVERHEAD: NORMAL VS. BURST ATTACK

Metric	Normal Operation	Attack Scenario
CPU Utilization (%)	1.42%	1.04%
Memory Usage (MB)	163.84 MB	163.90 MB
Avg. Processing Latency	3.14 ms	3.48 ms (est.)

displayed in, respectively, Fig. 7, 8, and 9. The computation overhead is summarized in Table III both for normal and burst-attack conditions.

F. Advanced Scenarios and Scalability Evaluation

To further assess the robustness and real-world deployability of the proposed behavior-aware access control framework, additional experiments were conducted focusing on large-scale deployments and adversarial attack strategies. These scenarios address the ability of the gateway to 1) scale to hundreds of concurrently connected IoT devices and 2) defend against "smart" attackers that attempt to evade anomaly detection by mimicking normal request rates.

1) Simulation scaling with 500+ virtual devices: A large-scale simulation was executed to evaluate whether the gateway can sustain high device counts while maintaining low latency and stable performance. In this experiment, 520 virtual IoT devices were registered and authenticated at the gateway. Each device issued two legitimate requests to an authorized endpoint, resulting in a total of 1,040 requests processed during the run.

The objective of this scenario was to measure scalability-related indicators, including successful request handling, access denials, processing latency, and gateway resource usage.

TABLE IV. SCALING RESULTS FOR 500+ VIRTUAL DEVICES

Metric	Value
Virtual devices	520
Requests per device	2
Total requests	1,040
Successful responses (200)	1,040
Denied responses (403)	0
Other errors	0
Average latency (ms)	1.68
Latency samples	1,040
Gateway CPU usage (after run)	0%
Gateway memory usage (MB)	164.44

TABLE V. DETECTION RATES FOR MIMICRY ATTACKS (NORMAL RATE, UNAUTHORIZED ENDPOINTS)

Attack Scenario	Endpoint Targeted	Requests Sent	Detected (403)	Rate (%)
Mimicry: normal rate	/admin/train	25	25	100
Mimicry: normal rate	/admin/delete	25	25	100
Mimicry: normal rate	/internal	25	25	100
Mimicry: normal rate	/system	25	25	100
Mimicry: normal rate	/root	25	25	100

All requests in this scenario corresponded to normal behavior and authorized access patterns.

Table IV illustrating the results demonstrate that the gateway successfully handled over 500 concurrent devices and more than 1,000 requests without any access denials or errors. The average processing latency remained below 2 ms, indicating that the integration of behavioral monitoring and lightweight ML does not introduce significant overhead even at scale. These findings confirm the suitability of the proposed framework for deployment in medium- to large-scale IoT environments.

2) Adversarial evaluation mimicry attacks: In addition to the scalability, the framework was tested in mimicry attacks listed in Table V. This is a more sophisticated attack where an adversary mimics the attributes of normal traffic patterns (including frequency and timing), and establishes a connection with an untrusted endpoint.

In this scenario, attackers issued requests at a normal rate (one request every 0.5 seconds) but targeted forbidden administrative and internal endpoints. The detection objective was to verify whether the gateway could prevent unauthorized access even when behavioral patterns appeared benign. Importantly, these attacks are expected to be blocked primarily by access control policy enforcement, rather than anomaly score thresholds alone.

In all the evaluated mimicry scenarios, all the unauthorised requests made against the endpoints selected for the Gateway were denied by the Gateway with a HTTP 403 response. It's important to keep in mind that this result is based on the endpoints tested, the number of requests, and the access-

control policies applied. It demonstrates that in such cases, the attackers were not able to avoid the system simply by mimicking the legitimate request timing for access to explicitly forbidden resources.

3) Discussion: The advanced evaluation confirms that the proposed system scales effectively to hundreds of devices while maintaining low latency and minimal resource consumption. Furthermore, the mimicry attack results demonstrate resilience against sophisticated attackers who attempt to evade anomaly detection by imitating legitimate behavior. Together, these findings reinforce the practicality of deploying the framework in real-world IoT environments, where both scalability and adversarial robustness are essential requirements.

VII. Challenges and Techniques

A. Resource Constraints at the Gateway

IoT gateways have limited computational and memory resources in comparison to cloud servers. Their tasks include authentication, protocol translation, traffic forwarding, and enforcement of security policies, often under strict real-time constraints. Deploying ML-based security mechanisms at this level is therefore challenging, particularly when low latency is required [2].

To address these constraints, the proposed framework employs lightweight feature extraction and a computationally efficient anomaly detection model. The behavioral features are derived from simple statistical properties of request activity (e.g., request frequency and endpoint access patterns) rather than deep packet inspection or payload analysis. These features are consistent with those available in widely used IoT security datasets, such as the ToN_IoT dataset, which emphasizes telemetry and flow-level characteristics suitable for resource-constrained environments.

Furthermore, the system avoids frequent retraining and reliance on large labeled datasets. The Random Forest model operates in a supervised manner and is trained primarily on normal behavior derived from the ToN_IoT dataset, resulting in lower memory usage and reduced computational overhead. This design ensures that continuous security monitoring does not interfere with the gateway's primary networking functions.

B. Model Accuracy vs. Computational Overhead

One trade-off inherent in the IoT security systems is the trade-off between the accuracy of detection and the cost of computation. As high-accuracy models can be trained with the aid of complex deep learning models, this is usually not feasible in resource constrained gateways because the training and inference is expensive [9].

The proposed system is focused on feasible deployability, whereby it is chosen to select a lightweight anomaly detection model that is both accurate enough with an acceptable overhead. More so, feature scaling and model inference are carried out efficiently and real-time assessment of each device request is possible. Even though lightweight models might not be able to measure very sophisticated attack patterns, the experimental findings suggest that they could be used to identify common post-authentication anomalies with low latency. This trade-off is reasonable in the case of enforcement of the gateway level, where the response timeliness is of great importance and the marginal accuracy gains are not always so important.

C. Dealing with Concept Drift in Behavior of Devices

The IoT device behavior can change with the passage of time because of the change in application requirement, update of firmware or the condition of use. Changes like a concept drift may make the behavior occurred earlier in normalcy to seem drastic hence producing false positives [3].

To overcome the above issue, the proposed framework will accommodate the re-training process periodically with verified benign behavior gathered at the gateway. The system can adapt to the gradual changes in behavior, while still recognizing the sudden or extreme changes. But retraining needs to be carefully done so that the animals do not become poisoned or learn bad habits. Therefore, on the borderline cases, it is recommended to manage them temporarily by giving restricted access instead of blocking them immediately, minimizing the operational impact of false alarms, while maintaining the security.

D. Security and Robustness Issues

Although the suggested framework improves the security of IoT by enforcing behavior-based access control, various issues of robustness still exist. The enemies can strive to imitate normal behavior so as to escape detection of anomaly or they could progressively adjust their behavior so as not to cause thresholds. Also, there are such attacks as poisoning against the training data, which may adversely affect the performance of the models [6].

To overcome such issues, the system restricts training of models to the times when they are confirmed to be operating benignly and limits administrative access to training functionality. The selection of the behavioral features is aimed at capturing many dimensions of interaction with the device and thus increasing the challenge to the attacker to avoid detection across all dimensions at the same time. Even though improved adversarial attacks are a research problem, the suggested design offers a viable and robust base to the defense of security by the gateway [38] [26].

E. The Security vs. Usability Trade-off

The results show that there is a clear compromise between recall (to avoid missing attacks) and precision (to affect operational usability). The proposed Random Forest implementation focuses on the anomaly-class recall and obtained unrounded anomaly-class recall at 0.9986. A critical infrastructure or industrial Internet of Things application may be able to tolerate this preference due to the potential cost of a missed attack being greater than the cost of a false alarm. However, the design decision should be carefully interpreted as high false positive rate may have an impact on the legitimate IoT operation.

Many samples were considered to be anomalous given the operating threshold and are reflected in the number of false positives, which was 22,470. In reality, this can result in a delay to legitimate device communication, alert volume or temporary denial of access for devices with abnormal but benign activity, for example sensors that do not read smoothly, but read bursts during a change in environmental conditions. Therefore, a

production deployment must have a staged response: low-confidence anomalies should be blocked in restricted mode and only high-confidence malicious behavior should be blocked immediately. Other mitigation measures are device specific baselines, calibration of thresholds, review by an administrator for 'borderline' cases, whitelisting of safety-critical endpoints, and retraining periodically with verified benign data.

VIII. Proposed Behavior-Aware Access Control Framework And Evaluation

A. Authorization Workflow Behavior-Aware

The suggested behavior-sensitive authorization protocol is a continuation of conventional IoT access control implemented with the evaluation of continuous behavior analysis following authentication. After a successful authentication at the gateway, a device receives an initial access level, and it is put under the close supervision. A request by the device made after a certain behavior is utilized to create a new behavioral profile indicative of real-time operational properties [39].

Behavioral characteristics are derived out of recent interactions with the device and sent to the lightweight anomaly detection model. A scoring on the resulting anomaly is the extent of deviation of the learned normal behavior. This score is then instantly translated to the authorization decision, which allows the system to dynamically change access privileges. The proposed workflow will establish a continuous reassessment of trust in the authorization pipeline instead of presumed forever by embedding behavior analysis within it [4].

The interaction between the IoT device, gateway and the ML model are illustrated in Fig. 10 for real-time authorization. A REST request is received by the gateway and behavioral features are extracted from the request and sent to the trained ML model. The anomaly score is then sent back to the gateway and correlated with an authorization decision, such as allow, restrict, or block.

B. Enforcement of Real-time Access Control

One of the distinguishing characteristics of the proposed system is the possibility to impose decisions in access control in real time. The proposed framework is also able to implement changes in authorization immediately an authorization change is detected unlike conventional intrusion detection systems that only send out alerts without implementing changes in authorization.

Devices with normal behavior are allowed to access all the authorized resources whereas suspicious devices are redirected into restricted access mode with restricted endpoint accessibility and reduced request rates. Gates ban the access of devices that are classified as malicious at the gateway. Such a graduated response mechanism reduces the impact of legitimate devices as well as facilitates the quick containment of the potentially dangerous devices.

The enforcement logic is directly applied at the gateway whereby every request is assessed and governed without involving any services. This design also contributes greatly to speed of response and allows to mitigate threat proactively in IoT environments that have latency constraints [13].

C. Comparative Study to Traditional Approaches

The proposed framework has significantly better resilience to post-authentication attacks compared to the conventional identity-based access control systems. The existing authorization models cannot react to the corrupted machines that are authorized after a credential is proven, but the proposed system can constantly modify the access privileges in accordance with observed behavior [31].

Unlike cloud-based intrusion detection systems, a lightweight ML implementation in the gateway level minimizes the communication overhead as well as conserves the privacy of the data and yields quicker enforcement measures. Even though deep learning-based methods can be used to perform better when detection rates are high and in a controlled environment, their computational costs and lack of connection with access control make them impractical in enforcing security in real-time IoT settings [29].

In general, the suggested framework balances security effectiveness, computational efficiency, and deployability, which is why it is more applicable to a real-world Internet of Things than most of the existing solutions.

D. Comparative Performance Analysis and Baseline Comparison

For test purposes, the framework is expanded to include the traditional identity-based access control and cloud-based deep learning IDS systems. The proposed framework is envisioned as a gateway level ML solution, hence a representative lightweight ML IDS baseline is included. This baseline is edge, gateway IDS approaches that are lightweight, do local ML based anomaly detection but not directly link detection results to dynamic access control enforcement.

The proposed gateway-based framework using random forest algorithm is summarized in Table VI with the traditional RBAC/ABAC, the lightweight ML IDS baselines deployed by the gateway and the deep learning IDS approaches deployed in the cloud.

1) Analysis of improvements: The proposed system has achieved 99.86% or 0.9986 Recall (Anomaly Class) instead of 100%. This number gets rounded only to 1.00 if rounding to 2 decimal places. The framework is more advanced than the traditional RBAC/ABAC systems by introducing post-authentication monitoring and dynamic enforcement. Its primary benefit, over its gateway-deployed lightweight ML IDS counterpart, is not only local detection, but the direct mapping of anomaly results to access-control decisions. The framework offers local enforcement with lower latency, which is not possible with cloud-based deep learning IDS based approaches, as it takes an average of 3.14 ms to process the data in the evaluated environment. But, the comparison needs to be taken with a pinch of salt, as datasets and metrics vary between studies, as do deployment assumptions.

IX. Discussion

A. Security Implications

Behavior-conscious ML combined with access control can greatly improve the security of the IoT through post-authentication threats. The proposed system mitigates the

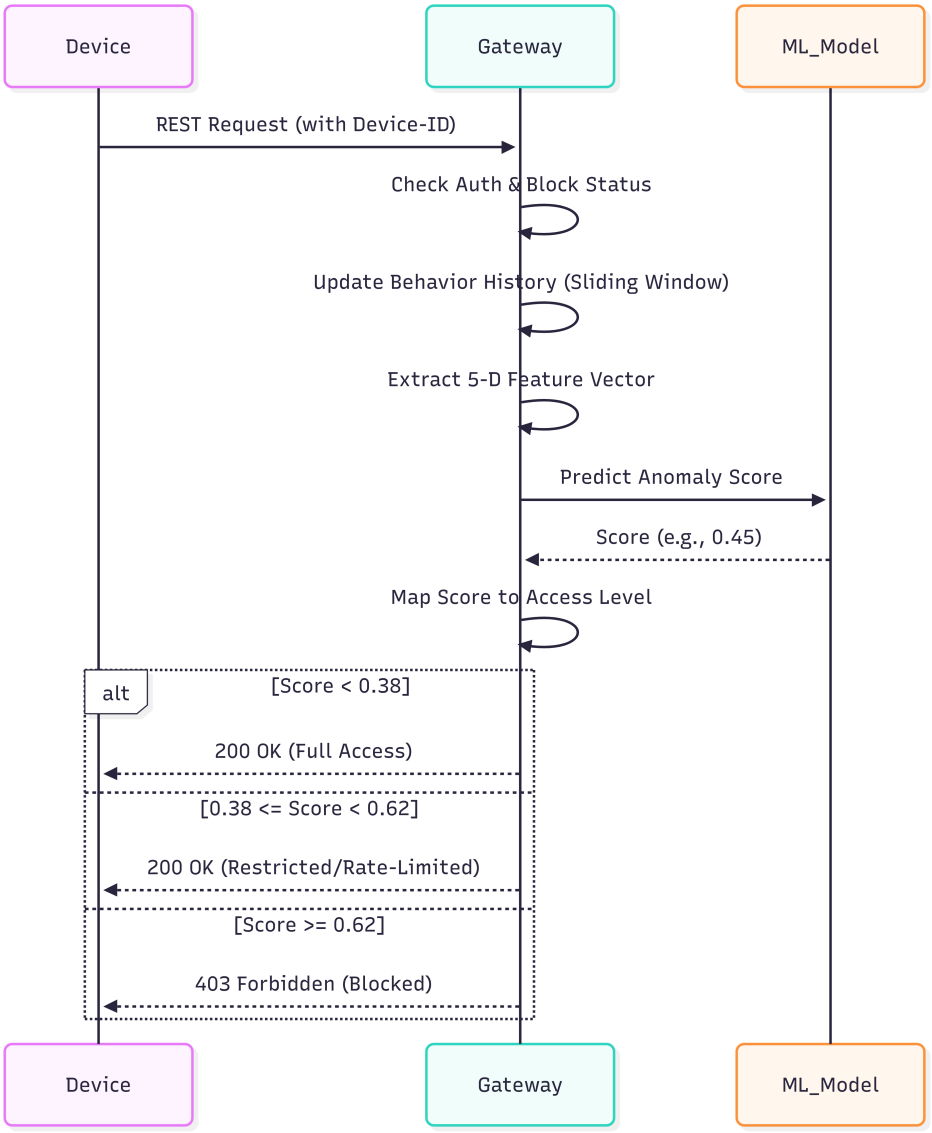


Fig. 10. Real-time request processing and dynamic authorization mapping.

TABLE VI. BASELINE COMPARISON: PROPOSED FRAMEWORK VS. TRADITIONAL, LIGHTWEIGHT ML, AND CLOUD-BASED IDS MODELS

Feature/Metric	**Traditional RBAC/ABAC**	**Gateway-Deployed Lightweight ML IDS Baseline**	**Cloud-Based DL IDS**	**Proposed Framework**
Post-authentication monitoring	None	Passive anomaly monitoring	Passive anomaly monitoring	**Active monitoring and enforcement**
Deployment level	Gateway/local policy	Gateway/edge	Cloud or edge-cloud	**Gateway/edge**
Model type	No ML	Lightweight classifiers such as Decision Tree, Random Forest, or k-NN	CNN, LSTM, DBN, or hybrid DL	**Random Forest plus access-control policy engine**
Detection performance	N/A	High but dataset-dependent; usually evaluated as IDS accuracy/recall only	98.8%–99.5% reported in related work	**99.86% anomaly recall; rounded to 1.00 only at two decimal places**
Processing latency	Low, but no anomaly detection	Low to moderate depending on gateway hardware	High due to cloud transmission and heavier inference	**Very low average processing latency of 3.14 ms in the evaluated setting**
Access-control link	Static	Usually decoupled from access control	Decoupled from gateway authorization	**Tightly coupled with allow, restrict, and block decisions**
Main limitation	Cannot detect post-authentication compromise	Detects anomalies but usually requires separate/manual enforcement	High computational cost, bandwidth use, and privacy concerns	**Higher false positives; dataset-specific validation; requires retraining for concept drift**

threat of insider attacks, credential theft, and compromised nodes because it constantly assesses the conduct of devices [30].

Supervised anomaly detection can help identify malicious behavior that statistically resembles learned attack profiles; however, its ability to detect entirely unseen attacks depends on the representativeness of the training data and the selected behavioral features.

B. Practical Deployability

In a practical sense, the suggested framework is highly applicable to the implementation in the real-world IoT gateways. The ML model and feature extraction are lightweight, which guarantees low computational overhead, and the gateway-centric design makes it easy to integrate the new system with the existing IoT system.

The framework is not dependent on modifying individual IoT devices, which means that it can be used with heterogeneous and legacy deployments. Moreover, gateway-based local processing promotes the concept of privacy by not relaying sensitive behavioral information to other servers [27].

C. Comparison to Cloud-Based IDS Solutions

The cloud-based intrusion detection systems are scalable and have a great computational capability but are affected by the latency, bandwidth, and privacy issues. Conversely, the solution suggested at the gateway level is capable of detecting and enforcing on-the-fly without any connection to the cloud.

Although cloud-based systems can be the right choice in the large scale analytics and long term threat intelligence, the suggested structure fits better the real-time protection and access control implementation. The combination of the gateway-level enforcement and a cloud-based analysis might also be further improved in the future to enhance the security of the deployments [9] [32].

D. Limitations and Threats to Validity

The envisioned framework illustrates a promising behavior-aware enforcement at gateway level, but some limitations must be pointed out. The evaluation is given based on the ToN-IoT benchmark dataset, which is the first one. ToN-IoT also offers heterogeneous IoT telemetry and attack traffic, but does not necessarily contain all real-world deployments, device types, device firmware behaviors or attack strategies. Thus, results reported should be viewed as findings of effectiveness for the evaluated data set and scenarios, and not as general assurances of performance.

Second, the system is based on a fixed five dimensional statistical feature representation (mean, stddev, range, median, non-zero ratio). This representation facilitates lightweight gateway deployment but might lack some protocol-level, temporal, or semantic aspects that are necessary to identify some advanced attacks from atypical device activity. The number of false positives seen in the experiment suggests further refinement of the features or calibration of the devices for production deployment.

Thirdly, the attack concept and patterns change in real IoT environments. Firmware updates, environment changes, new applications or changes in user behavior may alter normal device behaviour. If not retrained at regular intervals, and if the threshold is not adjusted, the model can make false-positive or false-negative errors. Lastly, the experimental gateway was tested in a lab setting, and the framework should be further tested on physical IoT testbeds, multiple deployment of gateways, and extended observation periods.

X. Conclusion and Future Work

This study presented a behavior-aware access control framework by embedding Supervised Random Forest learning into the access control logic for an IoT gateway. The study also demonstrated that the lightweight five-dimensional statistical summaries of the ToN-IoT benchmark dataset can benefit the low-latency edge anomaly detection.The study further confirmed that the five-dimensional statistical summaries of the ToN-IoT benchmark dataset, including mean, standard deviation, range, median, and non-zero ratio, can help with low latency anomaly detection at the edge. The experimental evaluation gave a Recall of 0.9986 (99.86%) on a test set of 120,000 samples and 125 false negatives in 91,294 attack instances, which is an anomaly-class recall. The results should thus be reported as “near-complete detection” for the scenarios considered instead of “perfect” or “universal” attack blocking. The framework illustrates the potential for Zero Trust inspired enforcement on resource limited gateways, without the need for cloud-based deep learning services, however, the false positive results, dependency on a training set, fixed feature representation, and concept drift problems need to be further validated before being deployed in real-world production systems.

The next study will focus on evaluating the proposed framework in a larger scale of IoT, where hundreds or thousands of devices are involved. Scalability issues, including not only higher behavioral diversity but also heavier load on the gateways will be explored to guarantee the stability of performance. Moreover, such approaches as federated learning or online learning might enable multiple gateways to effectively improve their models without needing to share raw data. These could improve flexibility to new threats and privacy as well as minimize centralization of dependencies [40]. Also, the suggested behavior-sensitive authorization model is consistent with the concepts of Zero Trust security. The future could also further integrate with the Zero Trust architectures, where ongoing authentication and least-privilege access are used throughout all the layers of the IoT ecosystem.

Author Contribution

All authors have equally contributed. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency fo Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia under the [GRANT No. KFU263248].

Data Availability Statement

The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments

The authors wish to express their gratitude to the Deanship of Scientific Research, Vice Presidency fo Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia. We would like to acknowledge the anonymous reviewers who made great contributions with their brilliant scholarly intuitive comments and sagacious recommendations to improve the quality and clarity of this study.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in internet of things: The road ahead," *Computer Networks*, 2025.

[2] M. A. Amanullah, R. A. A. Habeeb, F. H. Nasaruddin, A. Gani, E. Ahmed, A. S. M. Nainar, and M. Imran, "Deep learning and big data technologies for iot security," *Computer Communications*, 2020.

[3] E. Gyamfi and A. Jurcut, "Intrusion detection in internet of things systems: A review on design approaches leveraging multi-access edge computing, machine learning, and datasets," *Sensors*, vol. 22, no. 10, 2022.

[4] A. Verma and V. Ranga, "Machine learning based intrusion detection systems for iot applications," *arXiv preprint arXiv:2302.12452*, 2023.

[5] "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Information Security and Applications*, vol. 50, 2025.

[6] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "Ddos in the iot: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2025.

[7] M. Wazzan, D. Algazzawi, O. Bamasaq, A. Albeshri, and L. Cheng, "Internet of things botnet detection approaches: Analysis and recommendations for future research," *Applied Sciences*, vol. 11, no. 12, 2021.

[8] A. L. Buczak and E. Guven, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, 2025.

[9] A. Awajan, "A novel deep learning-based intrusion detection system for iot networks," *Computers*, vol. 12, no. 2, 2023.

[10] N. Pandey and P. K. Mishra, "Detection of ddos attack in iot traffic using ensemble machine learning techniques," *Networks and Heterogeneous Media*, vol. 18, no. 4, pp. 1393–1409, 2023.

[11] A. Zanella, N. Bui, A. Castellani, and L. Vangelista, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.

[12] M. M. Rashid, J. Kamruzzaman, M. M. Hassan, T. Imam, and S. Gordon, "Cyberattacks detection in iot-based smart city applications using machine learning techniques," *International Journal of Environmental Research and Public Health*, vol. 17, 2020.

[13] X. Kong and W. Yuhan, "Edge computing for internet of everything: A survey," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2022.

[14] Noshina et al, "Securing the internet of things in artificial intelligence era: A comprehensive survey," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2024.

[15] A. Bilal, Q. Riaz, M. Zeeshan, H. Tahir *et al.*, "Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using unsw-nb15 data-set," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, 2021.

[16] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for internet of things," *RMIT Repository*, 2024, posted 2024-11-02.

[17] N. Moustafa, B. Turnbull, and K.-K. R. Choo, "An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things," *IEEE Internet of Things Journal*, vol. 6, pp. 4815–4830, 2024.

[18] S. Alhasan, G. Abdul-Salaam, and Y. M. Missah, "Hybrid network intrusion detection systems: A systematic review," *Preprint*, 2024, january 2024.

[19] P. Cardoso, "Internet of things: Security challenges and solutions," *Preprint*, 2022, june 2022.

[20] E. Hodo, X. Bellekens, A. W. Hamilton, and P.-L. Dubouilh, "Threat analysis of iot networks using artificial neural network intrusion detection system," 2016.

[21] A. Sivanathan, D. Sherratt, H. H. Gharakheili, and V. Sivaraman, "Low-cost flow-based security solutions for smart-home iot devices," 2016.

[22] B. Kitchenham, O. P. Brereton, D. Budgen, and M. Turner, "Systematic literature reviews in software engineering–a systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, 2025.

[23] Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto, and K. Sakurai, "Machine learning-based iot-botnet attack detection with sequential architecture," *Sensors*, vol. 20, 2020.

[24] L. Xiao, X. Wan, and D. Wu, "Iot security techniques based on machine learning: How do iot devices use ai to enhance security?" *IEEE Signal Processing Magazine*, January 2020, published 19 January 2020.

[25] C. Shao, X. Du, J. Yu, and J. Chen, "Cluster-based improved isolation forest," *Entropy*, vol. 24, no. 5, 2022.

[26] G. Kalleshappa and B. Savadatti, "Network traffic analysis through deep learning for detection of an army of bots in health iot network," *International Journal of Pervasive Computing and Communications*, 2022, ahead-of-print.

[27] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of things security and forensics: Challenges and opportunities," *Future Generation Computer Systems*, vol. 78, pp. 544–546, 2025.

[28] N. B. Karankar and A. Seth, "A comprehensive survey on internet of things security: Challenges and solutions," pp. 711–728, 2023.

[29] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, 2021, 37 pages; also available as arXiv:2104.07914 [eess.SP].

[30] M. Alazab, S. Priya, P. M, and P. K. Reddy, "Federated learning for cybersecurity: Concepts, challenges and future directions," *IEEE Transactions on Industrial Informatics*, October 2021.

[31] E. Abdulrahman, S. Alshehri, and A. Cherif, "Blockchain-based access control for the internet of things: A survey," 2021.

[32] J. Oh, S. Yu, J. Lee, S. Son, M. Kim, and Y. Park, "A secure and lightweight authentication protocol for iot-based smart homes," *Sensors*, vol. 21, 2021.

[33] A. Singh and K. Chatterjee, "Trust based access control model for securing electronic healthcare system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 2, pp. 783–794, 2025.

[34] L. Zhou and H.-C. Chao, "Multimedia traffic security architecture for the internet of things," *IEEE Network*, vol. 25, no. 3, pp. 35–40, 2025.

[35] S. Suganth and D. Usha, "A survey of intrusion detection system in iot devices," *International Journal of Advanced Research*, vol. 6, no. 6, pp. 23–30, 2025.

[36] B. R. Kikissagbe and M. Adda, "Machine learning-based intrusion detection methods in iot systems: A comprehensive review," *Electronics*, vol. 13, no. 18, 2024.

[37] S. B. Sharma and A. K. Bairwa, "Leveraging ai for intrusion detection in iot ecosystems: A comprehensive study," *IEEE Access*, vol. 13, pp. 66 290–66 317, 2025.

[38] S. Krishnapriya and S. Singh, "A comprehensive survey on advanced persistent threat (apt) detection techniques," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 1–10, 2024.

[39] A. Diro, S. Kaisar, A. V. Vasilakos, A. Anwar, A. Nasirian, and G. Olani, "Anomaly detection for space information networks: A survey of challenges, techniques, and future directions," *Computers & Security*, 2024, open Access.

[40] M. Abomhara and G. M. Køien, "Cyber security and the internet of things: Vulnerabilities, threats, intruders and attacks," *Journal of Cyber Security and Mobility*, vol. 4, no. 1, pp. 65–88, 2025.

A New Drilling Rate of Penetration Prediction Model by Particle Swarm Optimization and Gradient Boosting Regression

Faris Aiman Jamaluddin[1], Marina Yusoff[2], Diva Kurnianingtyas[3], Mohamad Taufik Mohd Salledud-din[4]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI),
Universiti Teknologi MARA (UiTM), Shah Alam, 40450, Selangor, Malaysia[1]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI),
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam, 40450, Selangor, Malaysia[2]
Faculty of Computer Science, Universitas Brawijaya, Malang, 65145, Indonesia[2,3]
PETRONAS Research Sdn Bhd, Kawasan Institusi Bangi, Bandar Baru Bangi, Selangor[4]

***Abstract*—The oil and gas industry continuously evolves to enhance operational efficiency and productivity while minimizing costs and environmental impact. Among the critical aspects of oil and gas operations, drilling efficiency is a key factor in accessing underground hydrocarbon reservoirs. Traditional machine learning models and current regression models have shown limitations in accurately modelling the rate of penetration due to the high nonlinearity of data. This project focuses on the rate of penetration prediction for drilling optimization. This study proposed a new drilling rate of penetration prediction model with the embedding of particle swarm optimization in a gradient boosting regression method. A solution representation of the particle is introduced as a hyperparameter strategy to explore the optimal parameter for predicting drilling datasets. Extensive experiments were carried out using two splitting strategies, One for All and All for One, across the drilling wells. The proposed GBR+PSO hybrid model achieved a mean absolute error of 1.205, representing a reduction of approximately 89.68% compared to the best-performing baseline model, K-Nearest Neighbor with the One for All splitting strategy, which achieved a mean absolute error of 11.68. The hybrid solution could enhance drilling ROP predictions, advancing the drilling rate of penetration strategy. It has the potential to support the development of autonomous drilling optimization, thus contributing to more efficient, reliable, and cost-effective drilling rate of penetration strategies in future drilling operations.**

Keywords*—Drilling; gradient boosting regression; machine learning; rate of penetration; particle swarm optimization

I. Introduction

The oil and gas industry, along with its related sectors, is continually seeking innovative and competitive ways to enhance operational efficiency and productivity [1] [2]. The objective is primarily to minimize costs and reduce the environmental impact. Reaching this objective requires improving processes in oil and gas operations; one of them is the drilling process to estimate and adjust the penetration rate (ROP) [3] [4]. In recent years, demands on the oil and gas sector have continued to increase. However, the cost of offshore exploration and development is very high, especially for hydrocarbon production [5]. The ROP is a a key drilling performance indicator through subsurface formations, regardless of whether the drilling is targeting oil, gas, or any other substance. ROP plays a vital role, serving as a fundamental metric in drilling operations in various industries, primarily in oil and gas exploration [6], geothermal energy production [7], mineral exploration [8], and environmental monitoring [9]. In spite of the critical role of ROP prediction in drilling optimization, understanding how drilling variables influence ROP remains an open question in drilling [10][11][12]. ROP presents a highly nonlinear problem. A lot of efforts had been established as early as a traditional models as such Bourgoyne and Young model in 1974 [13] and until an advanced AI methods [14][17][18].

Despite these significant advancements, existing studies still require further improvement in optimization efficiency, adaptive feature engineering, computational complexity reduction, explainability, robustness against geological variability, and real-time adaptive learning capability for intelligent drilling applications. Although ML has become more popular in drilling engineering, there is still a lack of thorough studies that concentrate on ML and optimization methods for ROP prediction. Therefore, the main aim of this research is to improve the ROP prediction by using a hybrid approach of ML and optimization methods while utilizing the University of Stavanger Rate of Penetration (USROP) dataset as a main resource [19]. Key questions addressed include the necessity of incorporating ROP variables into ML models, the comparative effectiveness of the hybrid approach versus traditional models for ROP estimation, and the utilization of predictive models for optimization purposes. The research questions are:

- How to construct a new proposed model compared to existing machine learning methods in terms of MAE for predicting ROP in drilling operations?
- What is the effect of different training, testing, and validation splitting scenarios on the predictive performance of the proposed model across varying drilling well conditions?
- To what extent does the improved prediction accuracy of the proposed model translate into measurable gains in drilling efficiency, intelligent drilling optimization, and real-time operational decision-making?

This study proposes a hybrid data-driven method to predict ROP by integrating GBR and PSO. Our contributions in hybrid ROP prediction are:

- The study introduces a new GBR+PSO hybrid model for ROP prediction, demonstrating enhanced predictive performance with significantly lower MAE values than existing machine learning methods.
- A thorough training, testing, and validation incorporating MAE analysis and benchmarking, splitting scenario of the drilling well.
- The approach improves drilling efficiency by increasing a significant prediction accuracy across different conditions in intelligent drilling optimization and operational decision-making strategies.

The organization of the study is structured as follows: Section II presents the related work. Section III is dedicated to the preliminaries, which consists ROP optimization and prediction, score metrics and a proposed scenario. In Section IV, the explanation of the material and methods includes the proposed hybrid GBR+PSO method and performance measure. The results and performance evaluations are discussed in Section V and Section VI, respectively, and Section VII concludes the study.

II. Related Work

Recent advancements in ROP prediction have increasingly adopted artificial intelligence (AI), machine learning (ML), and deep learning (DL) approaches to improve drilling efficiency and predictive reliability [20]. Accurate ROP prediction remains challenging due to nonlinear drilling behavior, noisy sensor signals, heterogeneous geological formations, concept drift, and high-dimensional drilling parameters [21][22]. Inappropriate feature selection may introduce redundant variables, resulting in reduced prediction accuracy and poor model generalization [20]. Consequently, recent studies have emphasized intelligent feature selection, adaptive learning, and optimization strategies to enhance prediction performance and real-time drilling capability [21][22]. For instance, Attention Mechanism-enhanced Bidirectional Long Short-Term Memory (AM-BiLSTM) model integrated with Kalman filtering and mutual information analysis for feature selection, where dominant drilling parameters such as weight on bit (WOB), total depth, weight on hook, and flow pumps were identified as significant predictors [20]. Even though AM-BiLSTM model achieved significant improvements in Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) compared to Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and BiLSTM-based models, further enhancement is still required to strengthen feature dependency learning capability. In particular, improving adaptive feature representation and model generalization across varying drilling environments remains a critical challenge.

In another research, a hybrid framework combining Fuzzy C-Means, Dynamic Time Warping, Genetic Algorithm, and Extreme Learning Machine, where correlation analysis was employed for feature selection and GA was used to optimize the ELM model under different drilling conditions [12]. The GA-optimized ELM achieved the best prediction accuracy among comparative methods, highlighting the importance of drilling condition identification and optimization strategies for accurate ROP prediction. Furthermore, a hybrid online ROP prediction framework incorporating formation drillability (FD) sensing, Savitzky-Golay filtering, and incremental learning strategies, where the inclusion of FD as an additional predictive feature improved prediction accuracy by at least 19% compared to six existing methods [22]. The similar performance on the optimized LSTM achieved the best overall performance with prediction precision greater than 96.87%, minimum Average Relative Percentage Error (ARPE), MAPE, RMSE, and bias, as well as maximum R^2 performance [10].

III. Preliminaries

A. ROP Optimization Process and Prediction

ROP optimization is possible in numerous ways. Fig. 1 illustrates two general methods for this optimization process on ROP. The ROP model is utilized to determine the optimal drilling parameters for a specific optimization problem, such as minimizing drilling time, drilling MSE, or other objectives [19]. This is then used to establish standard drilling parameters as such of weight on bit or drill bit in rotary speed (RPM). Fig. 1(a) and Fig. 1(b) indicate the difference between continuous learning [20][19] and a reference well method [19]. As seen in Fig. 1(a), the model is able to be constructed in real-time using continuous learning. This necessitates training the model while drilling, which is more challenging to accomplish due to the computational requirements of the software and hardware. The first dataset is extremely small and skeletal at the beginning, which is another disadvantage that slows down the machine learning training process. There is a warm-up phase in which no model is used or a temporary model is employed. A reference well, as demonstrated in Fig. 1(b) is analogous to a drilled well. This enables users to construct models offline that need to be done only once, and is generally more easier compared than the continuous learning methods. In both methodologies, the optimum drilling parameters are computed using the ROP model and then applied in the drilling controls for parameters such as bit weight, RPM and mud flow [13]. The method would be based on local constraints and data availability. Best practices do not yet exist however, a hybrid approach is also feasible.

Effective ROP optimization depends on the development of a reliable ROP prediction model. Traditionally, both analytical and data-driven models were initiative and improvised to achieve this target [15]. In order to train the machine learning model or estimate model constants, both methods require reference datasets. Reality and the constructed models should have a smaller predicted error that would be references for the future drilling. This implies that the precision of a model can be utilized only for similar drilling such as tools, methods and depth. In modern drilling operations, continuous learning techniques have gained prominence [16][20]. Information to be referred, or trained, can be sourced from the local wells at a given time in a continuous learning process in parallel of the model creations.

B. Score Metrics

MAE is suggested as the key metric for ROP prediction.

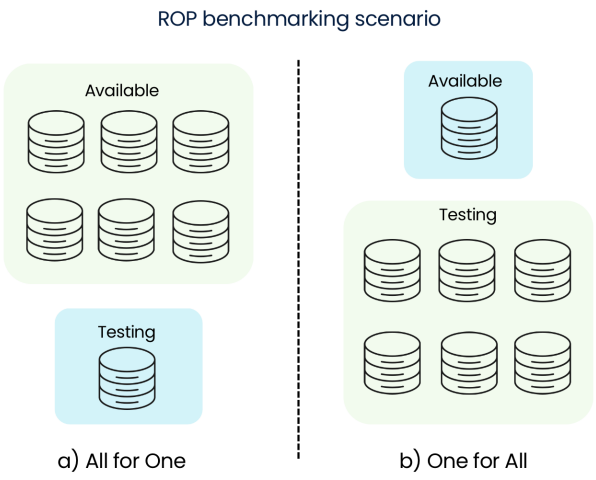


Fig. 1. ROP performance evaluation strategy.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{1}$$

where, n = number of samples, t = sample number consecutively, At = actual value at sample t, and Ft = forecasted value at sample t. The reason for this choice is that ROP modelling is mostly done for drilling time optimization, where the interest is in the cost per meter drilled. A specific error value, e.g., 10 m/h, will be of roughly the same significance to an operator whether the true value is 30 m/h or 80 m/h [19]. While MAE is widely used, other studies may employ alternative performance metrics such as the MAPE and the Weighted Mean Absolute Percentage Error (WMAPE). MAPE, defined as a percentage of the actual values used to estimate error, can be useful for relative comparisons of performance across ROP ranges. However, it becomes problematic when actual values approach zero. WMAPE, on the other hand, handles sample imbalance by weighting errors by the true values, making it a more robust alternative in highly heterogeneous datasets. The evaluation measure must be properly selected based on the specific purpose and data characteristics of the study in ROP prediction.

C. Proposed Scenario

In ROP prediction, three performance evaluation strategies are adapted from the previous work of Tunkiel et al. [19] to evaluate ROP prediction models. The scenarios are All for One (AFO) and One for All (OFA). AFO, as illustrated in Fig. 1(a), in which one complete well is set aside for testing and all but one well are available for training and validation. This necessitates a final MAE score for all wells combined and enables seven different runs, with varying wells set aside for testing as cross-validation. Fig. 1(b) shows the One for All (OFA) situation, where only one well is used to estimate, and the rest are estimated by this model, as suggested by the final situation. This creates seven train/test iterations as cross-validation, just like the second scenario. Regarding iteration independence, in both the AFO and OFA strategies, each iteration is treated as an independent entity. This demands that the algorithms designed must be able to work with data from a single iteration without being influenced by preceding iterations. This habit would make it possible to guarantee that the models remain objective and do not inadvertently incorporate irrelevant information that could undermine their precision.

IV. Materials and Techniques

This section elaborates on the materials, data sources, and research methodologies described. The proposed approach captures the steps to see the performance of the enhancement of GBR+PSO models on ROP prediction optimization. The approach includes data preparation, pre-processing, construction of the proposed methods, and evaluation. In addition, we introduced a particle representation or solution mapping for the PSO. Fig. 2 demonstrates the flow of the proposed methodology. This process begins with the data extraction and pre-processing stage, where raw data is cleaned of outliers, missing values are imputed, overlapping data are removed, and the data are sorted. The pre-processing stage is rigorously carried out to ensure the equality and consistency of datasets, thereby ensuring that high-quality data are gathered and creating a strong model. The subsequent feature selection yields an optimal dataset containing the most descriptive predictors of the ROP model. The essence of the methodology lies in the predictive modelling achieved using machine learning techniques.

Three regressors namely, AdaBoost, Gradient Boosting, and K-Nearest Neighbours are assessed for performance across multiple algorithmic paradigms. Hyperparameter optimization made possible through PSO is applied to optimize the performance of selected models. PSO is a population-based stochastic optimization procedure well suited to navigating complicated, high-dimensional search spaces, making it suitable to determine optimal parameter configurations.

Model validation involves using performance measures such as MAE and R^2, which provide numerical measures of predictive accuracy and model fitness. The convergence analysis of the PSO algorithm is also conducted to ensure the stability and reliability of hyperparameter tuning. The design system consists of two approaches, namely AFO and OFA, enabling the comparison of model performance across different training and test paradigms.

The availability in both the training and the testing phases leads to a comprehensive analysis of the models. The training process involves tuning the selected machine learning models to the preprocessed data, while the testing process assesses the generalization performance of the models on unseen data. The integration of a combination of robust preprocessing, advanced optimization techniques, and rigorous validation presents an all-encompassing and methodological solution to ROP prediction, which could be beneficial in generating useful knowledge for enhancing drilling efficiency and optimizing resource recovery in shale reservoirs. Strategic comparison between AFO and OFA methods enables the diversification of model training and testing performance under varying operational conditions. Detailed steps are elaborated in the following sub-sections.

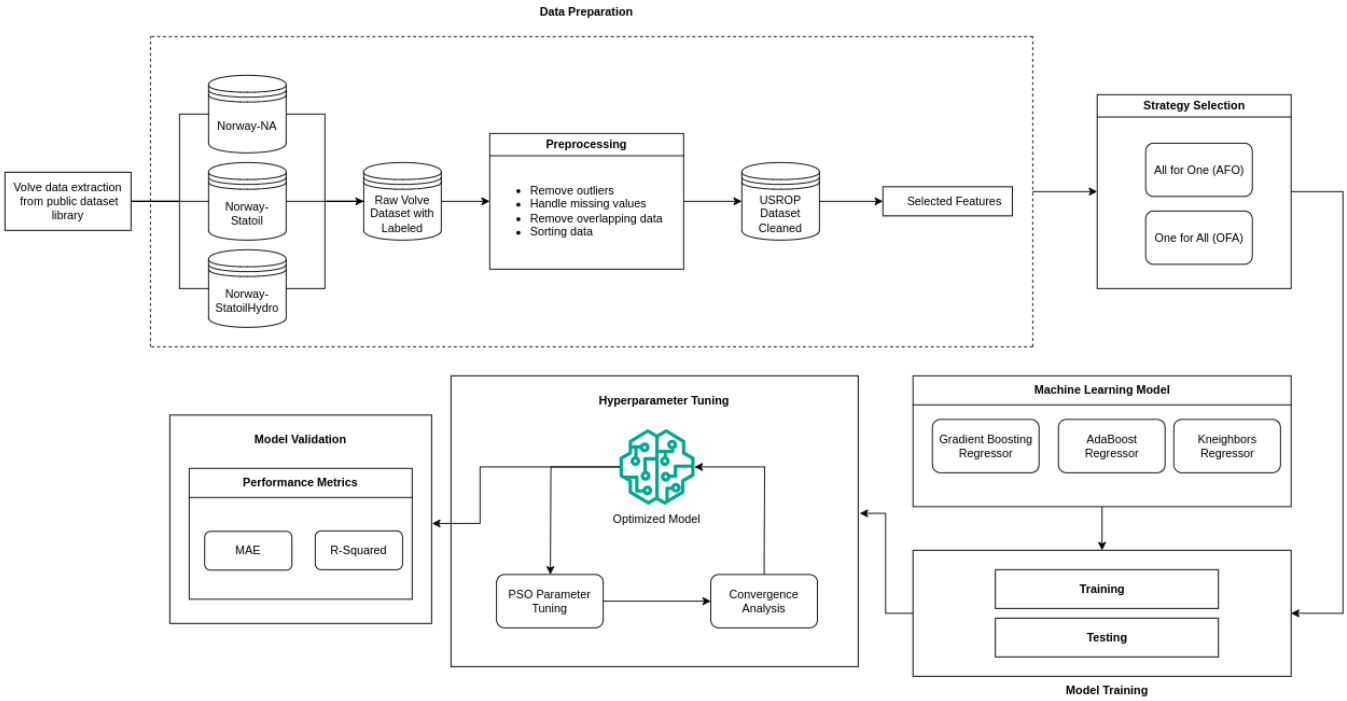


Fig. 2. Flow of methodology.

A. Data Preparation

This study utilized the same precise data preparation as that of the USROP reference paper. Experiments in numerous drilling areas are facilitated by the Volve database's large-scale real-time drilling data. Meanwhile, the logs differ in features, contain missing values, and vary in quality. A good subset of Volve must address this problem. Seven wells in total were selected based on the logged attributes. Data quality and the logged common attributes were the selection criteria. The datasets were acquired from the USROP dataset, focusing on the Volve field. [19]. The attributes are selected: Measured Depth [m], Weight on Bit [kkgf], Average Standpipe Pressure [kPa], Average Surface Torque [kN.m], Rate of Penetration [m/h], Average Rotary Speed [rpm], Mud Flow In [L/min], Mud Density In [g/cm3], Diameter [mm], Average Hook-load [kg], Hole Depth (TVD) [m], USROP Gamma [gAPI]. Minimal processing was done to the attributes to preserve the original data. This is necessary as the drilling logs often contain erroneous, non-physical values. There may be sentinel values indicating no reading (typically -999), corrupted values from the mud-pulsing system, and others. Samples containing weights on bit values below 0 and above 35 were truncated. The same way rows with mud density in, mud flow in, and average surface torque values below zero were removed, as well as with ROP values above 100 and average standpipe pressure above 25,000. Meanwhile, the diameter refers to the nominal wellbore diameter. Forward and backward filling were used to fill small gaps in the data resulting from uneven logging frequency across different equipment.

Distribution of ROP values by frequency across five ranges shows a pronounced leftward skew. The concentration of data in the 1-20 m/h and 21-40 m/h ranges suggests that the majority of drilling operations experienced quite low penetration rates. This may indicate natural limitations imposed by the penetrated geological rocks, such as high rock strength or unfavourable drilling conditions. Alternatively, it can indicate suboptimal drilling operations or equipment inefficiencies that prevent higher ROP. The declining rate observed with increasing ROP intervals also indicates the challenge of achieving high penetration rates. The explanation for this trend could be rising depth, which typically leads to higher compressive stresses and poorer rock drillability. The diminishing returns of increasing weight on bit (WOB) or rotary speed at high ROP values could also be a factor for this trend. A detailed investigation of geological conditions, drilling parameters, and operating practices is critical to elucidate the underlying causes of this skewed ROP distribution and to identify potential means of enhancing drilling performance.

B. Feature Correlation Coefficient

This correlation heatmap illustrates the interrelations among various drilling parameters and their impact on ROP (Fig. 3). Colour intensity and direction (red positive, blue negative) indicate the strength and direction of the linear relationships among variables. Surprisingly, ROP also shows a moderate negative correlation with Measured Depth (-0.39) and Hole Depth (TVD) (-0.42), indicating a decrease in ROP with increasing depth, perhaps due to factors such as increasing formation compaction and pressure. Conversely, there is a moderate positive correlation between ROP and Mud Flow In (0.25), suggesting the potential for greater ROP at higher mud flow rates, possibly due to more effective cuttings transport and hole cleaning.

Even though this heatmap is useful as a first-level view

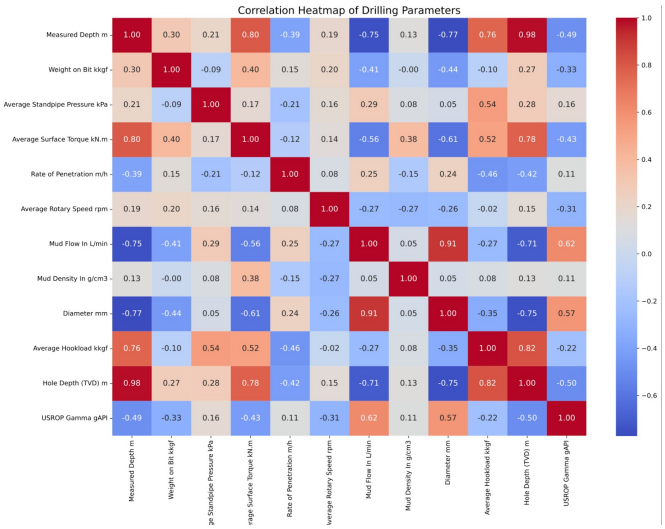


Fig. 3. Features correlation

TABLE I. PARTICLE REPRESENTATION FOR GBR TUNING

Variable	Hyperparameter range
X_1	`n_estimators`: 50–500
X_2	`learning_rate`: 0.01–0.3
X_3	`max_depth`: 3–10
X_4	`min_samples_split`: 2–20
X_5	`min_samples_leaf`: 1–20
X_6	`subsample`: 0.5–1.0
X_7	`max_features`: 0.1–1.0

of drilling parameter correlations with ROP, it has its own limitations. First, the correlations presented are specific to the dataset employed and may not extrapolate to other geological structures and drilling conditions. Second, the interpretation is restricted to linear relationships, whereas the actual interaction of parameters and ROP might be more complex and non-linear. Therefore, further study using new analysis techniques, such as non-linear regression modelling and machine learning algorithms, is needed to develop a larger, predictive ROP estimation model. Further consideration of business sense and operational feasibility will become crucial to demonstrate the actual, real-world practicability of the prediction model when implemented.

C. Proposed Solution

1) Solution representation: Following is the solution mapping or particle representation (see Table I):

whereas,

X1 to X7 are labels used to denote specific GBR parameters in a structured format.

R1 to R7 are shorthand notations corresponding to the ranges of parameters X1 to X7.

X1, R1: n_estimators: Range from 50 to 500.

X2, R2: learning_rate - Range from 0.01 to 0.3.

X3, R3: max_depth - Range from 3 to 10.

X4, R4: min_samples_split - Range from 2 to 20.

X5, R5: min_samples_leaf - Range from 1 to 20.

X6, R6: subsample - Range from 0.5 to 1.0.

X7, R7: max_features - Range from 0.1 to 1.0.

2) Fitness function: The fitness function was calculated using test data from previous references (Tunkiel et al., 2021a). The equation given below describes the fitness function used to calculate MAE, which is applied within the PSO algorithm.

$$\min\left(\frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|\right) \tag{2}$$

whereas,

leftmargin=1.15em

- n: Total number of data points (hours) for drilling speed.
- t: Counter that iterates through each data point, going from 1 to n. Keeps track of the specific hour analysed within the calculation.
- A_t: Actual drilling speed measured at a specific hour (t). A_1 would be the speed for the first hour, A_2 for the second, and so on.
- F_t: Predicted drilling speed for the same hour (t).
- $|A_t - F_t|$: Difference between the actual speed (A_t) and the predicted speed (F_t). It ignores the direction of the error (positive or negative).
- $\sum$: Sum of all the absolute differences between actual and predicted speeds for all the hours ($t = 1$ to $t = n$).
- $(1/n)$: Represents the average.

In this setup, MAE serves as the objective function which PSO aims to minimize. MAE is mainly used to measure the accuracy of the GBR model's predictions. Each particle in the PSO algorithm represents a potential set of GBR hyperparameters. The fitness of each particle is evaluated by training the GBR model with its corresponding hyperparameters and calculating the MAE on a validation dataset.

3) A proposed new hybrid GBR+PSO: In this study, a new hybrid GBR+PSO algorithm is proposed (see Algorithm 1). The following is a step-by-step of a new hybrid GBR+PSO algorithm. The PSO algorithm iteratively updates particles' positions and velocities based on their own best-known positions, their neighbours' best-known positions, and the global best-known position. This update process continues until the algorithm converges on the optimal hyperparameters that yield the lowest MAE, ensuring the most accurate ROP predictions. This dynamic interplay between MAE and PSO effectively guides the search for the optimal hyperparameters by utilizing PSO's exploration and exploitation capabilities to fine-tune the GBR model. Based on Barbosa et al. [23] has been shown few past studies have used the MAE as an objective function for ML optimization. In this PSO setup for GBR hyperparameter tuning, the particles' parameters (hyperparameters for the GBR model) and the objective function (MAE) are tightly linked:

- Particles' parameters (GBR hyperparensuree and train the GBR model.
- The MAE is computed based on the predictions of the GBR model with tcouldhose hyperparameters and serves as the objective function that PSO optimizes.
- The goal of PSO in this setup is to minimize the prediction error of the GBR model. MAE is a straightforward measure of average error and fits well as the objective function for this minimization problem.

Algorithm 1 A New Hybrid GBR+PSO

- Begin.
- Initialize the PSO algorithm for GBR hyperparameter tuning.
- Define the particle representation.
- Initialize the swarm with a population of particles, each representing a candidate set of GBR hyperparameters.
- Set the fitness function as MAE, calculated using Equation (2), where: leftmargin=1.5em,noitemsep,topsep=1pt
  - n: Total number of data points (hours) for drilling speed
  - t: Iteration counter
  - A_t: Actual drilling speed at time t
  - F_t: Predicted drilling speed at time t
- Evaluate the initial fitness of each particle by training the GBR model using its hyperparameters and calculating the MAE on a validation dataset.
- Update the velocity and position of each particle based on: leftmargin=1.5em,noitemsep,topsep=1pt
  - The particle's personal best position (pBest)
  - The global best position found by the swarm (gBest)
- Iterate until the stopping criterion is met (e.g., a fixed number of iterations for convergence to the lowest MAE).
- Select the optimal GBR hyperparameters corresponding to the particle with the lowest MAE.
- Train the final GBR model using the optimized hyperparameters.
- Measure the model training time.
- Predict the test dataset using the trained GBR model.
- Evaluate the final GBR model's performance using MAE.
- End.

GBR was selected for advanced hyperparameter tuning using PSO, with validation performed using the AFO splitting strategy . Unlike the initial phase, this phase utilized all relevant GBR parameters to achieve optimal model performance. The PSO setup, as outlined in Table III, was configured with 100 iterations and 100 particles, default values to ensure robust exploration of the hyperparameter space. The inertia weight was set to 0.5, while the cognitive and social coefficients were both set to 2.0. These parameters were chosen to balance exploration of new solutions and exploitation of known good solutions, ensuring the algorithm efficiently converges to the best hyperparameter configuration. The PSO algorithm works by iteratively adjusting the position of each particle based on its personal best solution and the global best solution found by the swarm. The inertia weight controls the influence of a particle's previous velocity, while the cognitive and social coefficients guide the particles toward self-improvement and collaboration, respectively. This approach ensures a comprehensive search of the hyperparameter space, minimizing the MAE and enhancing the predictive accuracy of the GBR model.

D. Performance Evaluation

This study has used two essential empirical measurements to evaluate and compare the effectiveness of GBR+PSO. By running this experiment and applying PSO to the GBR, the main goal of this study is to identify the optimal set of hyperparameters for the GBR, including the learning rate, the number of trees, and tree depth, that minimise prediction error while maximising overall model performance, as measured by the MAE metric. This optimization step is an important part of my experiment, as it is represented to significantly enhance the model's accuracy and robustness. Plus, these improvements will contribute to more precise and reliable ROP predictions and enable more efficient and informed decision-making in drilling operations. A proposed hybrid GBR+PSO was compared with GBR, ADR and KN models. In evaluating the performance of GBR+PSO, the hyperparameter ranges for GBR were carefully selected to ensure meaningful exploration and avoid impractical configurations. The lower and upper bounds for the hyperparameters were set as follows: `n_estimators` (50 to 500), `learning_rate` (0.01 to 0.3), `max_depth` (3 to 10), `min_samples_split` (2 to 20), `min_samples_leaf` (1 to 20), `subsample` (0.5 to 1.0), and `max_features` (0.1 to 1.0). These ranges were selected based on their significant influence on the performance of the GBR model. For instance, `n_estimators` control the number of boosting iterations, with higher values potentially improving accuracy at the cost of increased computational time. The `learning_rate` determines the contribution of each tree to the final model, with smaller values promoting more robust learning but requiring more iterations. The `max_depth` parameter limits the depth of each tree, preventing overfitting by controlling model complexity. Similarly, `min_samples_split` and `min_samples_leaf` balance underfitting and overfitting by regulating the minimum number of samples required to split a node and the minimum number of samples per leaf node, respectively. The `subsample` parameter introduces randomness by using a fraction of samples for fitting each tree, enhancing generalization. Finally, `max_features` controls the proportion of features considered for each split, reducing overfitting while maintaining predictive accuracy.

By leveraging PSO to fine-tune these hyperparameters within the specified bounds, the GBR model achieves an optimal balance between predictive accuracy and computational efficiency. The iterative nature of PSO ensures that the best hyperparameter configuration is selected based on the training data, resulting in a robust, high-performing model. This approach not only enhances the predictive capabilities of the GBR model but also provides a systematic and efficient method for hyperparameter optimization, making it a valuable contribution to the field of machine learning and predictive

modeling. In this implementation, PSO is used to fine-tune the hyperparameters of the GBR. The optimization process involves 100 iterations, especially during which particles (candidate solutions) explore the search space to minimize the MAE. Each particle will then adjust its position based on its previous experience (personal best) and the best-known solution found by the swarm (global best). The algorithm utilizes inertia weight to control how much a particle's previous velocity influences its current movement, balancing exploration and exploitation. These two coefficients guide the particles: the cognitive coefficient, encouraging self-improvement, and the social coefficient by promoting collaboration toward the best global solution. The parameter ranges are constrained by specific lower bounds ([50, 0.01, 3, 2, 1, 0.5, 0.1]) and upper bounds ([500, 0.3, 10, 20, 20, 1.0, 1.0]) to ensure meaningful exploration. These settings allow PSO to iteratively refine hyperparameter combinations, efficiently converging to an optimal GBR configuration.

The GBR is a machine learning algorithm that builds decision trees to minimize prediction error. In this implementation, the hyperparameters optimized by PSO should significantly influence GBR's performance as follows:

leftmargin=1.15em

- `n_estimators`: The number of boosting iterations (trees) in the model, ranging from 50 to 500. More iterations can improve accuracy but may increase computation time.
- `learning_rate`: Determines the contribution of each tree to the final model. A smaller value (e.g., 0.01 to 0.3) results in more robust learning but requires more iterations.
- `max_depth`: Limits the depth of each tree (3 to 10), controlling model complexity and preventing overfitting.
- `min_samples_split`: Minimum number of samples required to split a node (2 to 20), balancing underfitting and overfitting.
- `min_samples_leaf`: Minimum number of samples per leaf node (1 to 20), affecting tree size and generalization.
- `subsample`: Fraction of samples used for fitting each tree (0.5 to 1.0), introducing randomness for better generalization.
- `max_features`: The proportion of features considered for each split (0.1 to 1.0), which reduces overfitting while maintaining predictive accuracy.

By tuning these hyperparameters based on lower bound and upper bound, the GBR model is able to achieve an optimal balance between predictive accuracy and computational efficiency. Next, PSO ensures the selection of the best configuration based on the training data.

V. Computational Results

This section highlights the computational results of GBR, KN and ABR and the proposed GBR+PSO. The experiments were perform based on an appropriate parameters for each of the methods, GBR and ABR, utilizing the different learning rate of 0.1 and 0.01 meanwhile KN based on the number neibours. All experiments tested on AFO and OFA benchmark splitting strategy.

A. Computational Results Using GBR

GBR employed the OFA methodology with a learning rate tinkered to 0.01 and the same 80/20 train-test split. This experiment was therefore meant to investigate the lower learning rate's impact while targeting a particular drilling scenario. Similar to GBR-EXP1,GBR-EXP3, GBR-EXP4 has shown a promising learning curve. With little tweak on the parameter, we can clearly see the gap between the training and test error showing that to some extent the model is generalizing well. The training error goes down smoothly, which means it is learning from the data. However, the train error also goes down in the beginning and then flattens. This would then mean that, in learning, there is an optimum size of training beyond which the generalization performance does not increase significantly. It appears that the optimal training size is around 60,000. Beyond this point, the model starts to overfit, and the test MAE starts to increase. This may be due to various factors involving data complexity, limitations of model architecture, or noise in data. Further analysis, such as hyperparameter tuning and feature engineering, could potentially improve the model's performance and further reduce the test error.

B. Computational Results Using KNR

In KNN-EXP1, the AFO method has been used with n_neighbors = 3. This setting reduces the neighbourhood size, thus should be allowing the model to capture more local fine details. The experiment uses an 80/20 train-test split to examine the model's behavior when trained on most of the dataset and validated on a smaller portion. Training MAE just remains very low since the model manages to easily fit to the data being trained from. On the other hand, test MAE shows its height relatively throughout the process. This large difference between the training and testing MAE indicates underfitting criteria of the model.

C. Computational Results Using ABR

The learning curves for ABR experiments from ABR-EXP1 to ABR-EXP4, all reflect the overfitting pattern. Whereas the training error decreases rapidly, indicating that the model fits the training data well, either the test error saturates or increases. This also reflects that the model memorizes the training data rather than learning the underlying pattern. This is further supported by the overfitting behaviour evident from the violin plot, since the MAE test values in the various experiments are widely distributed, with noticeably large variability in performance. This could mean that the current ABR model architecture, with its default hyperparameters, is overly simplistic for handling such a large dataset range of 120,000 to 200,000 samples. The model's complexity may not be sufficient to learn the complex relationships in such a large dataset. Hence, it overfits the training data and generalises very poorly with high prediction errors on unseen data.

D. Computational Results Using GBR+PSO

In assessing the impact of particle representation on a new hybrid GBR+PSO performance, the results GBR+PSO obtained MAE: 1.205, which indicate good MAE value and good optimization process and R^2: 0.982 indicates goodness-of-fit measure for linear regression models. From the results, it demonstrates that the hyperparameter tuning procedure, which was driven by PSO, has successfully discovered an optimal set of hyperparameters for the GBR model. This is demonstrated by the algorithm's iterative exploration of various hyperparameter combinations, which resulted in a "best new configuration" with improved MAE values at each step. In the PSO algorithm for hyperparameter tuning, some important parameters guide the search process. This parameter refers to cognitive factor (C1) and social factor (C2). The cognitive factor parameter, which is set to 2.0 by default, gives an idea of how much a particle is influenced by its own pBest to continue the exploration of good areas previously found. On the other hand, the social factor C2 = 2.0 will relate to the degree of attraction by a particle for the gBest discovered by all the swarm. That enhances collaboration or sharing of information between particles. Along with the inertia weight which is set at 0.5 to maintain a balance between exploration and exploitation the described parameters enable this swarm to dive into the space of hyperparameters. The PSO algorithm explored 100 iterations with a swarm of 100 particles evaluating many possible configurations or possible feasible solutions. Table II demonstrates the selected hyperparameter configurations of the Configuration 1 and Configuration 2.

TABLE II. THE SELECTED HYPERPARAMETER CONFIGURATIONS

Parameters	Configuration 1	Configuration 2
n_estimators	95	425
learning_rate	0.166916	0.132280
max_depth	3	6
min_samples_split	17	14
min_samples_leaf	8	9
subsample	0.716358	0.533660

Configuration 1 and Configuration 2 obtains MAE of about 4.178 and 1.205, respectively. The algorithm successfully identified a new best configuration in the second iteration, significantly reducing the MAE from 4.178 to 1.205. This improvement highlights the effectiveness of the PSO algorithm in optimizing hyperparameters, leveraging the interplay between C1, C2, and inertia weight to find a high-performing model. The best fitness (lowest MAE) achieved during the hyperparameter tuning process is 1.205, which corresponds to Configuration 2 in the training output demonstrating the practical benefits of using PSO for hyperparameter tuning.

Fig. 4 shows the "Predicted vs. Actual ROP" plot of GBR+PSO experiment provides a compelling visual representation of the model's performance, corroborating the findings from the learning curve and the hyperparameter tuning process. The strong positive correlation between predicted and actual ROP values, evident in the tight clustering of points around the diagonal line, demonstrates the model's ability to accurately capture the underlying relationships in the data. The low MAE which is 1.205 and of about R^2 of 0.982 values further support this observation. While the plot reveals a strong

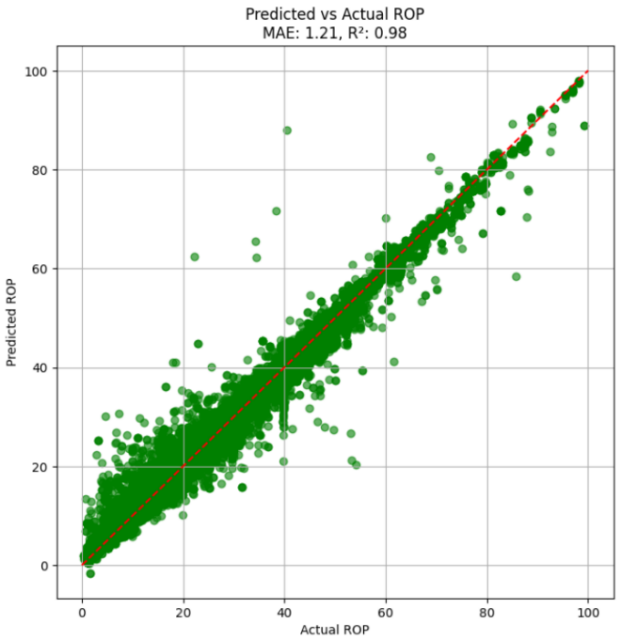


Fig. 4. Predicted versus actual ROP for the best GBR+PSO experiment.

overall correlation, it also provides valuable insights for further refinement. By analyzing the distribution of points and identifying areas where deviations from the diagonal line, we can pinpoint specific scenarios or data regions, where the model's predictions might be less accurate. This plot gives insight to guide further hyperparameter tuning, feature engineering, or data collection efforts to enhance model performance and improve the accuracy and reliability of ROP predictions.

VI. DISCUSSION

The experimental results presented in the tables provide a comprehensive comparison of various machine learning models, including GBR, Adaboost, KNR, and GBR+PSO. The performance of these models is evaluated using the MAE metric, which is particularly suitable for predicting the ROP in drilling operations. MAE measures the average absolute difference between predicted and actual values, providing a clear, interpretable metric of model accuracy. For ROP prediction, where even small deviations can have significant operational implications, MAE is an appropriate metric because it directly quantifies the average error in predicted ROP values, ensuring the model's predictions are reliable and actionable. The performance metrics for different models of the ROP prediction is demonstrated in Table III.

The superior performance of the GBR+PSO model, with an MAE of 1.205, highlights the effectiveness of PSO in fine-tuning hyperparameters. PSO's ability to efficiently explore the hyperparameter space and balance exploration with exploitation allows it to identify configurations that minimize prediction errors [24][25]. This is particularly important for ROP prediction, where the relationship between input parameters includes weight on bit, standpipe pressure, and rotary speed and ROP is complex and non-linear. The learning curve analysis for GBR+PSO would likely show a steady decrease in training and validation errors as the number of iterations

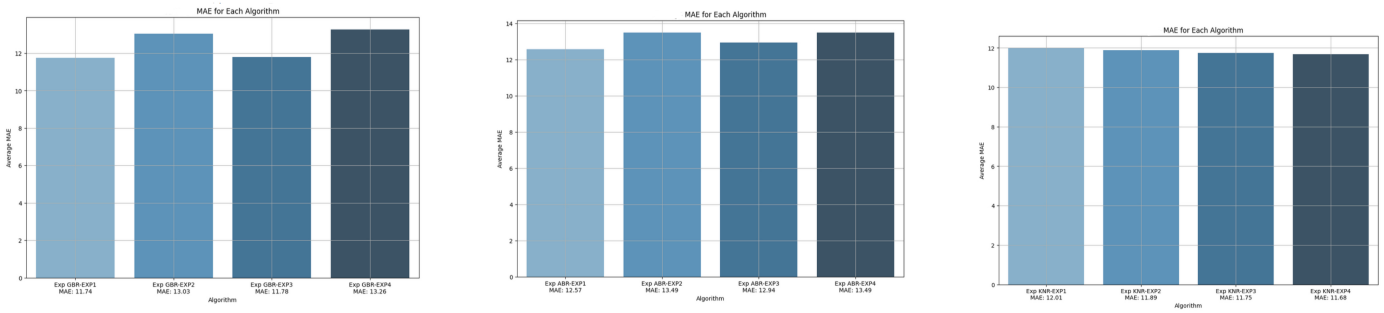


Fig. 5. MAE score for the standalone experiments: (a) GBR, (b) ABR, and (c) KNR.

TABLE III. PERFORMANCE METRICS FOR DIFFERENT MODEL CLASSIFIERS.

Method and Experiment ID	Benchmark Splitting Strategy	Parameter	MAE
GBR-EXP1	AFO	Learning Rate: 0.1	11.74
GBR-EXP2	AFO	Learning Rate: 0.01	13.03
GBR-EXP3	OFA	Learning Rate: 0.1	11.78
GBR-EXP4	OFA	Learning Rate: 0.01	13.26
ABR-EXP1	AFO	Learning Rate: 1.0	12.57
ABR-EXP2	AFO	Learning Rate: 0.01	13.49
ABR-EXP3	OFA	Learning Rate: 1.0	12.94
ABR-EXP4	OFA	Learning Rate: 0.01	13.49
KNR-EXP1	AFO	$n_neighbors = 3$	12.01
KNR-EXP2	AFO	$n_neighbors = 5$	11.89
KNR-EXP3	OFA	$n_neighbors = 3$	11.75
KNR-EXP4	OFA	$n_neighbors = 5$	11.68
GBR+PSO (Proposed Method)	AFO	Parameter range based on Table II	1.205

increases, indicating that the model is learning effectively without overfitting. This behavior is critical for ensuring that the model generalizes well to unseen data, which is essential for real-world drilling applications.

To further validate the model's performance, predict vs. actual plots are highly useful. These plots provide a visual comparison between the predicted and actual ROP values, allowing for a direct assessment of the model's accuracy. For a well-performing model like GBR+PSO, the points in the predict vs. actual plot should closely align with the 45-degree line, indicating minimal deviation between predictions and ground truth. This visualization confirms the model's accuracy and helps identify any systematic errors or biases in the predictions, which can be addressed in subsequent iterations of model refinement.

In summary, using MAE as a performance measure [20], along with learning curves, predict vs. actual plots, and violin plots, creates a strong way to assess and confirm the reliability of ROP prediction models. The GBR+PSO model's exceptional performance, as evidenced by its low MAE, illustrates the value of advanced optimization techniques and thorough model validation. By leveraging these tools, researchers and practitioners can ensure that their models not only achieve high accuracy but also exhibit consistent and reliable behavior across a wide range of drilling conditions, ultimately leading to more efficient and safer drilling operations (see Fig. 5).

VII. CONCLUSION

The research under this study has framed the potential of hybrid optimization techniques in improving the ROP prediction in drilling operations. This study has created a plan to make ROP predictions more accurate by mixing machine learning algorithms with traditional optimization methods, while also dealing with the challenges of drilling in areas that contain hydrocarbons. The results indicated that the methodologies introduced can deliver substantial improvements in drilling efficiency, enabling reductions in operational costs and increased productivity in the oil and gas industry. This work also underlined the importance of data-driven approaches in modern drilling practices. Given the increasing demands for efficiency and cost-effectiveness within the industry, this capability to predict ROP with good accuracy will become relevant. The findings of the research indicate that the developed hybrid GBR+PSO model would be very useful for drilling engineers in driving data-informed decisions powered by real-time and predictive analytics. This development does not just guarantee the operational success but also contributes to achieving those goals the industry has laid out concerning minimum environmental impact with maximum resource extraction. The objective of the research was to assess the overall effect of the suggested methodologies on drilling performance. The implemented hybrid optimization approach, when executed, resulted in a measurable increase in ROP, thereby increasing the overall productivity of the drilling operation.

This study is quite important because it shows that the practical benefits of adopting advanced predictive models are tangible in the field. The positive results associated with this objective are indicative of further research and development in the field of drilling optimization. Since the industry is in continuous evolution, advanced technologies and methodologies must be integrated not only to stay competitive but also to address the challenges posed by increasingly complex drilling environments. Despite these promising results, the study has certain limitations. The model was developed and validated using historical drilling datasets from a limited number of wells, which may not fully capture the variability of diverse geological formations and operational conditions encountered across the broader industry. Building on these findings, future work should focus on real-time field validation and deployment, testing the GBR+PSO model on live drilling rigs with real-time data streaming to evaluate its practical performance, latency, and reliability under actual field conditions, rather than solely on historical datasets. Further research could also explore integration with other optimization algorithms, hybridizing GBR with other metaheuristic algorithms such as Ant Colony Optimization, Grey Wolf Optimizer, or Whale Optimization Algorithm, to benchmark against PSO and identify the most robust optimization strategy. Additionally, future

studies should explore embedding the hybrid GBR+PSO model into autonomous or semi-autonomous drilling control systems to enable fully automated, closed-loop optimization of drilling parameters.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Faris Aiman b Jamaluddin: Writing – original draft, formal analysis, data curation. Marina Yusoff: Writing – review and editing, visualization, validation, supervision, methodology, and conceptualization. Diva Kurnianingtyas – validation, editing. Mohamad Taufik Mohd Salledud-din – data understanding, validation.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no competing interests that could have appeared to influence the work reported in this study.

ACKNOWLEDGMENT

The authors gratefully acknowledge support from the Universiti Teknologi MARA (UiTM), Malaysia, and the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Petronas Research Sdn.Bhd for knowledge transfer and financial support provided to this research project.

REFERENCES

[1] J. Li, L. Stamford, and A. Gallego-Schmid, "Full environmental life cycle costing analysis of repurposing onshore abandoned oil and gas wells for geothermal power generation," Applied Thermal Engineering, p. 130469, 2026.

[2] X. Yu and G. Ren, "A data-driven approach to achieve low-carbon building energy optimization by using BIM technology," International Journal of Advanced Computer Science and Applications, vol. 16, no. 8, 2025.

[3] Y. Chen, P. Chen, X. Dai, J. Gong, and K. Lu, "Prediction of rate of penetration in oil and gas well drilling: A hybrid data-driven approach with multi-stage data preprocessing," *Engineering Applications of Artificial Intelligence*,vol. 163, p. 113085, 2026.

[4] Y. Song, Z. Song, J. Yang, L. Wei, and J. Tang, "Enhancing energy efficiency and sustainability in offshore drilling through real-time multi-objective optimization: Considering lag effects and formation variability," *Reliability Engineering & System Safety*, vol. 261, p. 111138, 2025.

[5] S. Davoodi, M. Al-Shargabi, D. A. Wood, and M. Mehrad, "Advancement of artificial intelligence applications in hydrocarbon well drilling technology: A review," *Applied Soft Computing*, vol. 176, p. 113129, 2025.

[6] Z. W. Cai, R. L. Zhao, Z. M. Huang, Y. P. Qiao, Q. Y. Yue, C. L. Li, et al., "Research on the development trends of measurement while drilling (MWD) technology in oil and gas drilling," *Petroleum Science*, 2025.

[7] M. Y. Amer, S. K. Salem, M. S. Farahat, and A. M. Salem, "Reducing drilling cost of geothermal wells by optimizing drilling operations: Cost effective study," *Journal of Unconventional Resources*, vol. 7, p. 100196, 2025.

[8] Y. Zhang, L. Yu, L. Yang, Z. Hu, and Y. Liu, "Data-driven framework for predicting rate of penetration in deepwater granitic formations: A marine engineering geology perspective with comprehensive model interpretability," *Journal of Engineering Geology*, vol. 351, p. 108039, 2025.

[9] P. Chen, "Advancements and future outlook of safety monitoring, inspection and assessment technologies for oil and gas pipeline networks," *Journal of Pipeline Science and Engineering*, p. 100267, 2025.

[10] M. Naderi, A. Bagheri, and M. N. Khujin, "Hybrid LSTM-Desirability Concept for Drilling Rate of Penetration (ROP) Prediction and Optimization," Array, p. 100900, 2026.

[11] M. Bizhani and E. Kuru, "Towards drilling rate of penetration prediction: Bayesian neural networks for uncertainty quantification," *Journal of Petroleum Science and Engineering*, vol. 219, p. 111068, 2022.

[12] X. Yang, M. Wu, C. Lu, W. Li, L. Chen, and S. Du, "Prediction of rate of penetration based on drilling conditions identification for drilling process," *Neurocomputing*, vol. 579, p. 127439, 2024.

[13] A. Sauki, P. N. F. M. Khamaruddin, S. Irawan, I. Kinif, S. Ridha, S. A. Ali, and M. A. Ali, "Development of a modified Bourgoyne and Young model for predicting drilling rate," Journal of Petroleum Science and Engineering, vol. 205, p. 108994, 2021.

[14] R. Alharbi, N. Alageel, M. Alsayil, R. Alharbi, and A. A. Alhakamy, "Prediction of oil production through linear regression model and big data tools," International Journal of Advanced Computer Science and Applications, vol. 13, no. 12, 2022.

[15] A. Sharma, T. Burak, R. Nygaard, E. Hoel, T. Kristiansen, and M. Welmer, "Hybrid ROP modeling: Combining analytical and data-driven approaches for drilling," Geoenergy Science and Engineering, vol. 251, p. 213877, 2025.

[16] Yehia, T., Gasser, M., Ebaid, H., Meehan, N., & Okoroafor, E. R. (2024). Comparative analysis of machine learning techniques for predicting drilling rate of penetration (ROP) in geothermal wells: A case study of FORGE site. Geothermics, 121, 103028.

[17] R. Zhang, Z. Zhu, Z. Yan, T. Pan, X. Song, G. Li, et al., "EHTGNN: An Explainable Hybrid Temporal Graph Neural Network for robust rate of penetration prediction in drilling operations," Geoenergy Science and Engineering, vol. 259, p. 214347, 2026.

[18] R. H. Allawi, W. J. Al-Mudhafar, M. A. Abbas, and D. A. Wood, "Leveraging boosting machine learning for drilling rate of penetration (ROP) prediction based on drilling and petrophysical parameters," *Artificial Intelligence in Geosciences*, vol. 6, no. 1, p. 100121, 2025.

[19] A. T. Tunkiel, D. Sui, and T. Wiktorski, "Reference dataset for rate of penetration benchmarking," Journal of Petroleum Science and Engineering, vol. 196, p. 108069, 2021.

[20] Q. Yuan, M. He, Z. Chen, M. Liu, and X. Chen, "A real-time prediction method for rate of penetration sequence in offshore deep wells drilling based on attention mechanism-enhanced BiLSTM model," *Ocean Engineering*, vol. 325, p. 120820, 2025.

[21] D. Rezki, L.-H. Mouss, A. Baaziz, and T. Bentrcia, "Adaptive prediction of Rate of Penetration while oil-well drilling: A Hoeffding tree based approach," Engineering Applications of Artificial Intelligence, vol. 143, p. 111465, 2025.

[22] C. Gan, Y. Wang, W.-H. Cao, K.-Z. Liu, and M. Wu, "Real-time formation drillability sensing-based hybrid online prediction method for the rate of penetration (ROP) and its industrial application for drilling processes," *Control Engineering Practice*, vol. 164, p. 106487, 2025.

[23] L. F. F. M. Barbosa, A. Nascimento, M. H. Mathias, and J. A. de Carvalho, "Machine learning methods applied to drilling rate of penetration prediction and optimization - A review," Journal of Petroleum Science and Engineering, vol. 183, p. 106332, 2019, doi: 10.1016/j.petrol.2019.106332.

[24] M. Yusoff, D. Ehsan, M. Y. Sharif, and M. T. M. Sallehud-din, "Topology approach for crude oil price forecasting of particle swarm optimization and long short-term memory," International Journal of Advanced Computer Science & Applications, vol. 15, no. 1, 2024.

[25] M. Daviran, A. Maghsoudi, and R. Ghezelbash, "Optimized AI-MPM: Application of PSO for tuning the hyperparameters of SVM and RF algorithms," Computers & Geosciences, vol. 195, p. 105785, 2025.

An Experimental Evaluation of Deep Learning Networks for Automated Breast Cancer Detection

Partha Chakraborty*, Umme Aiman Jannat
Department of Computer Science and Engineering, Comilla University, Cumilla-3506, Bangladesh

***Abstract*—Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, where early and accurate diagnosis plays a vital role in improving survival rates. Recent advancements in deep learning have demonstrated significant potential in automating the analysis of medical images for cancer detection. This study presents a comprehensive comparative analysis of convolutional neural network (CNN)–based deep learning models for breast cancer classification using ultrasound and mammography images. Multiple architectures, including a baseline CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, were evaluated to classify breast lesions as benign or malignant. Experimental results reveal notable performance differences across imaging modalities. For ultrasound images, AlexNet achieved the highest accuracy of 89%, while DenseNet and the baseline CNN achieved 88% and 85%, respectively. In contrast, mammography-based classification yielded significantly higher performance, with the baseline CNN outperforming deeper architectures in terms of accuracy and F1-score, achieving 97%. The findings demonstrate that model complexity does not necessarily guarantee superior performance and that properly designed shallow CNNs can effectively outperform deeper networks on high-quality mammographic data. This study highlights the potential of deep learning–based computer-aided diagnosis systems to support radiologists in the early detection of breast cancer.**

***Keywords*—Breast cancer; deep learning; convolutional neural networks; mammography; ultrasound; medical image classification**

I. Introduction

Breast cancer is a major cause of death for women between the ages of 20 and 59. Breast cancer is the most common cancer in women worldwide. Bangladesh has one of the highest incidences of breast cancer in Asia. In Bangladesh, breast cancer is the most prevalent cancer among women aged 15 to 44 years, with an incidence of approximately 19.3 per 100,000 population compared to other cancer types. Early detection remains the most effective strategy to reduce mortality. Women under 40 are advised to perform regular breast self-examinations, while those over 40 are advised to undergo routine mammographic screening for early detection. Reducing death rates and increasing treatment outcomes depend heavily on early and precise identification. Despite their effectiveness, traditional diagnostic techniques are often limited by subjectivity and variability, making consistent and trustworthy diagnoses difficult. These constraints may be addressed by the revolutionary potential of machine learning (ML), which leverages enormous datasets and processing capacity to improve diagnostic accuracy and efficiency of diagnosis. Breast cancer happens when cells in the breast start growing in an abnormal and uncontrolled way, forming a lump called a tumor. If it is not treated in time, the cancer can spread from the breast to other parts of the body, such as the bones, liver, or lungs. It can affect both women and men, but it is far more common in women.

There are also early or precancerous conditions, often called precancerous or carcinoma in situ, in which abnormal cells are found in the milk ducts or lobules but have not yet spread into nearby tissues. Invasive breast cancer, on the other hand, begins to grow into the surrounding normal breast tissue and can potentially spread to other parts of the body [1]. The most common types include ductal carcinoma and lobular carcinoma. Other less common forms include Paget's disease of the breast and inflammatory breast cancer [2].

A prospective cross-sectional study was conducted at Ahsania Mission Cancer and General Hospital between July 2023 and December 2023 to investigate the frequency and causes of delayed presentation among patients with newly-diagnosed breast cancer. The study aimed to identify the extent of diagnostic delays, explore the underlying reasons, and examine associations with socio-demographic factors. The research involved 242 participants, primarily from low-income, uneducated backgrounds, with the majority (52.06%) aged between 41 and 60 years. Data was collected through face-to-face interviews and medical record reviews after obtaining informed consent. Diagnostic delay was defined as a period of 90 days or more between the onset of symptoms and the initiation of medical treatment by expert physicians. Findings revealed that nearly half of the patients (46.28%) experienced delays exceeding three months, with an average delay duration of 5.18 months. Stage II breast cancer was the most common diagnosis (56.6%). A statistically significant association was found between diagnostic delay and patients' socio-economic status as well as the stage at which cancer was diagnosed [3]. To address the critical issue of delayed breast cancer diagnosis, this study aims to enhance the performance of existing diagnostic approaches by validating deep learning–based models using real-world ultrasound and mammography datasets collected from multiple hospitals. This study utilizes datasets from mammography and ultrasound imaging to investigate the application of machine learning algorithms in breast cancer detection. Numerous machine learning models, including CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, were utilized and evaluated using key performance indicators such as accuracy, F1 score, weighted average, and macro average. The results show that CNN-based models are very successful, with mammography datasets outperforming ultrasound in terms of performance. In particular, CNN achieved the best accuracy for mammograms at 97%, followed by ResNet101 at 95% and ResNet50 at 94%.

*Corresponding author

With an accuracy of 89%, AlexNet outperformed other models for ultrasound data.

II. Related Works

Breast cancer detection has been extensively studied using machine learning (ML) and deep learning (DL) techniques, with numerous studies focusing on optimizing classification performance and early detection. Over the past decade, algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression, Naïve Bayes, and ensemble methods have been widely applied to both tabular and imaging datasets, demonstrating varying levels of effectiveness based on the nature of the data and preprocessing techniques employed [8], [20], [21].

Bazazeh and Shubair [9] conducted one of the earlier comparative studies on breast cancer detection using supervised ML techniques, including RF, SVM, and Bayesian Networks (BN). Their analysis highlighted the high predictive capability of BN, achieving a recall and precision of 97.1% and 97.2%, respectively, while RF exhibited optimal receiver operating characteristic (ROC) performance of 99.9%. The study emphasized the importance of leveraging historical labelled data for predictive modelling in clinical applications, demonstrating that even traditional ML techniques can provide high accuracy when appropriately tuned. Similarly, Mohammed et al. [22] and Bhise et al. [19] reviewed various ML and DL approaches, confirming the consistent use of SVM, KNN, RF, Logistic Regression, and Naïve Bayes in breast cancer detection. These studies collectively highlighted the strengths and weaknesses of each algorithm: SVM performs well in high-dimensional feature spaces but is less efficient for large or noisy datasets; KNN is conceptually simple and robust but computationally intensive for large datasets due to repeated distance calculations; RF effectively handles high-dimensional, categorical, and missing data, although hyperparameter tuning is essential to avoid overfitting; Logistic Regression is effective for binary classification but assumes linearity between independent and dependent variables; and Naïve Bayes is simple and fast, but its assumption of feature independence and the zero-frequency problem can limit real-world applicability.

In the context of image-based datasets, Charan et al. [4] utilized a large Kaggle breast cancer image dataset comprising 7,858 samples across four magnification levels. They applied ML classifiers, including SVM, KNN, RF, Logistic Regression, and Naïve Bayes, with feature extraction performed using convolutional and pooling operations. While competitive performance was achieved, SVM faced scalability issues on large, imbalanced image datasets. Selvathi and Poornila [24] further demonstrated that careful selection of classifiers and feature extraction methods is crucial when combining deep learning networks with traditional ML algorithms to optimize predictive performance.

Sharma et al. [7], [16] applied RF, KNN, Naïve Bayes, Logistic Regression, and Decision Tree classifiers on the Wisconsin Diagnostic Breast Cancer dataset. KNN achieved the highest accuracy of 95.90%, indicating its robustness for structured tabular data. Moreover, the study demonstrated that incorporating Principal Component Analysis (PCA) for dimensionality reduction and appropriate data preprocessing significantly enhances SVM performance. Specifically, SVM accuracy increased from 65.78% without preprocessing to 98.02% after PCA and Standard Scaling, and achieved 97.36% on test subsets after fine-tuning grid search parameters. These results underscore the importance of preprocessing and hyperparameter optimization in improving classifier performance.

Joshi and Mehta [10] applied image processing techniques along with ML classifiers on a set of 209 mammography images from 50 patients, using train-test-validation splits of 70:20:10. Relevance Vector Machine (RVM) achieved 97% accuracy, demonstrating that probabilistic sparse models can be highly effective even with reduced feature sets. However, the study noted that RVM is less frequently applied to breast cancer datasets, having been primarily utilized for the detection of lymphoma and leukemia [11].

Khuriwal and Mishra [12] combined deep learning with traditional classifiers (KNN, SVM, Decision Tree, and RF) on the mini-MIAS Mammographic Database. They applied comprehensive preprocessing steps including digitization, noise and artifact removal, background suppression, and pectoral muscle removal. Among the classifiers, RF achieved the highest accuracy of 98.89% in differentiating fatty and dense tissues, while SVM remained limited to binary classification, reflecting its well-known constraint in multi-class or high-dimensional imaging scenarios. Kumar et al. [13] further extended deep learning approaches by training a Convolutional Neural Network (CNN) on 70% of the MIAS dataset (322 images) using stochastic gradient descent with momentum (SGDM). Despite subdividing the dataset into seven classes (six abnormal), the model achieved only 65% accuracy for normal-versus-abnormal classification, highlighting the challenge of multiclass prediction in medical imaging.

Xiao et al. [14] proposed an unsupervised deep learning approach, employing feature extraction to reduce image dimensionality, followed by classification using stacked autoencoder–SVM (SAE-SVM). This approach demonstrated that combining unsupervised feature learning with supervised classification can improve predictive capability while reducing computational overhead. Halim et al. [15] implemented parallel AI models for early breast cancer detection, achieving faster results of simultaneously using multiple models. However, the study noted an increase in computational complexity due to model parallelism.

Ahmed et al. [17] analyzed the Wisconsin Breast Cancer dataset, containing 699 records (458 benign, 241 malignant), applying SVM, KNN, Naïve Bayes, RF, and Logistic Regression. All classifiers achieved over 95% accuracy, confirming the robustness of these algorithms for traditional tabular datasets. Vinayak et al. [18] applied ML techniques on cytological data from UCI, where Artificial Neural Networks (ANN) achieved 97.37% accuracy. Ensemble methods such as stacking and voting also produced competitive results, emphasizing the value of combining multiple classifiers for reliable predictions.

Ranjan et al. [6] highlighted the effectiveness of ensemble methods, including Gradient Boosting, AdaBoost, and XGBoost, with RF achieving perfect accuracy (100%), illustrating the advantages of ensemble learning while acknowledging limitations in real-world medical imaging applications. Easttom et al. [5] evaluated traditional ML models using the R

programming framework, finding that SVM achieved 94.71% accuracy but was constrained in high-dimensional imaging contexts.

The recent trend is the adoption of deep learning-based architectures such as CNNs, which automatically extract features from images without extensive preprocessing. CNNs can effectively differentiate between malignant and benign tumors by filtering and learning important image parameters, offering flexibility and robustness in image-based analysis. Contemporary implementations, often using Keras as a backend, have demonstrated superior accuracy and efficiency in automated breast cancer detection, forming the foundation of modern diagnostic systems [26], [27].

Overall, the reviewed literature demonstrates that both traditional machine learning and deep learning approaches have achieved promising performance in breast cancer diagnosis [23], [25]. Conventional classifiers, such as SVM, KNN, Random Forest, Logistic Regression, and Naïve Bayes, have demonstrated high accuracy on structured datasets, particularly when combined with effective preprocessing, feature selection, and dimensionality reduction techniques. However, these methods largely rely on handcrafted features and are sensitive to noise, class imbalance, and variations across imaging modalities. Recent studies increasingly favor deep learning models, especially Convolutional Neural Networks (CNNs), due to their superior capability in automatically extracting features and learning representations from complex medical images. Despite these advancements, most existing works are limited to a single imaging modality—primarily mammography or ultrasound—and focus mainly on classification, with limited attention to cross-modal analysis or comprehensive performance evaluation across diverse architectures. Furthermore, comparative studies involving multiple deep learning models under a unified experimental framework remain scarce. These limitations highlight the need for a robust and modality-aware diagnostic framework that systematically evaluates state-of-the-art deep learning architectures on both ultrasound and mammography data, while emphasizing reliable performance metrics beyond accuracy alone. Addressing these research gaps can significantly enhance diagnostic confidence and clinical applicability in automated breast cancer detection systems.

III. Methodology

The overall architecture of the proposed system is illustrated in Fig. 1. The framework begins with dataset acquisition, followed by image collection and preprocessing to enhance image quality and standardize input dimensions. The preprocessed images are then split into training and testing subsets. Deep learning models are trained using the training dataset, while the testing dataset is used to evaluate generalization performance. Model effectiveness is assessed using performance validation metrics, training and validation loss curves, and visual inspection of classification outcomes. Ultimately, the system generates a binary classification output indicating whether a case is benign or malignant. Fig. 1 provides a high-level representation of the complete workflow, highlighting the interaction between data preprocessing, model training, and evaluation components within the proposed framework.

While Fig. 1 presents the generalized architecture of the proposed system, it does not explicitly illustrate how different

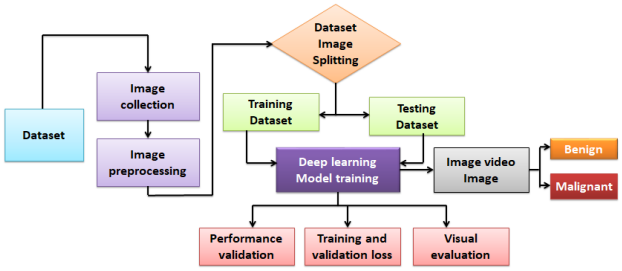


Fig. 1. System architecture

imaging modalities are handled independently. To address this, a more detailed experimental workflow is introduced in the next stage, focusing on modality-wise dataset splitting, model selection, and evaluation strategies for ultrasound and mammography images.

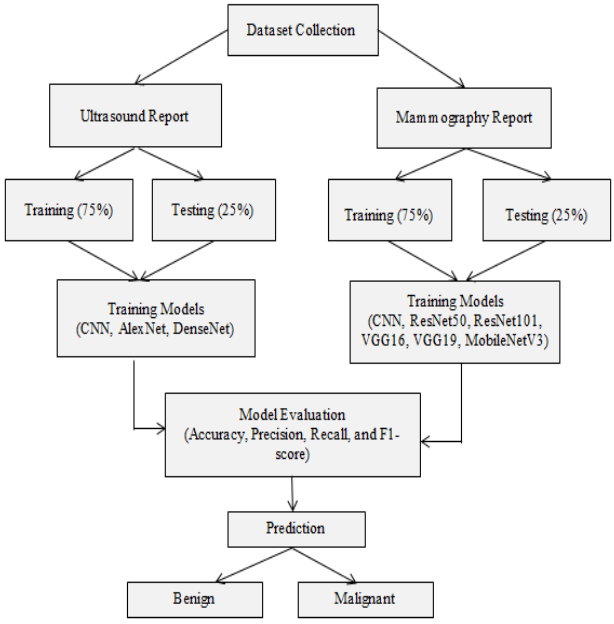


Fig. 2. Dataset-specific training and evaluation pipeline for ultrasound and mammography-based breast cancer classification.

The detailed experimental pipeline is illustrated in Fig. 2, outlining the dataset-specific model development process. The collected data are categorized into two imaging modalities: ultrasound reports and mammography reports. Each dataset is divided into 75% training and 25% testing subsets to ensure sufficient learning and reliable evaluation. For the ultrasound dataset, three models—CNN, AlexNet, and DenseNet—are trained to capture texture-based and low-contrast features commonly present in ultrasound images. In contrast, the mammography dataset is evaluated using CNN, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, enabling an in-depth comparison of shallow, deep, and lightweight architectures on high-resolution mammographic data.

A. Data Collection

The dataset used in this study was compiled from a combination of publicly available online repositories and offline clinical sources to ensure data diversity and reliability. The dataset consists of two primary imaging modalities: ultrasound and mammography. Ultrasound image data were collected from open-source platforms, including Kaggle, Hugging Face, and the UCI Machine Learning Repository.

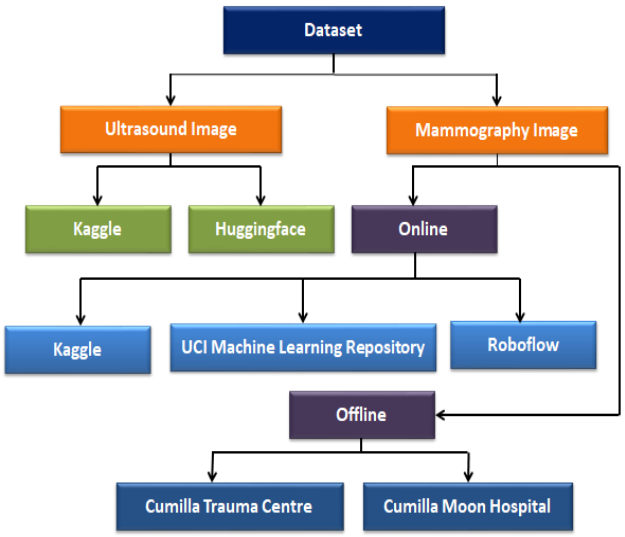


Fig. 3. Dataset collection process

Mammography images were obtained from online repositories, including the UCI Machine Learning Repository and Roboflow. To enhance clinical relevance, additional imaging data were collected from offline healthcare facilities, namely Cumilla Trauma Centre and Cumilla Moon Hospital. All offline data were acquired in compliance with institutional ethical guidelines, and patient-identifiable information was removed to maintain privacy and confidentiality. In this study, we have collected data from various sources, including Kaggle, UCI Machine Learning Repository, Roboflow, Hugging Face, and others.

Nearly 13k data have been collected. The aggregation of raw data from multiple online sources is used in the study (see Fig. 4). It showcases four publicly available platforms: Kaggle, UCI Dataset, Roboflow, and Hugging Face, each serving as an individual data source. All these sources are connected through a common pipeline that converges into a single repository labelled Raw Dataset. The study utilizes datasets collected from multiple public repositories and clinical sources, which may introduce variability due to differences in imaging protocols, acquisition devices, and annotation practices. Such source-specific biases and domain shifts can affect model generalization across different data distributions. While the proposed models demonstrate strong performance on the available datasets, further investigation using domain adaptation techniques and standardized datasets is necessary to improve robustness in real-world clinical settings. This diagram represents the initial stage of the data collection process, in which heterogeneous datasets from different platforms are combined into a unified raw dataset prior to further preprocessing, cleaning, and analysis.

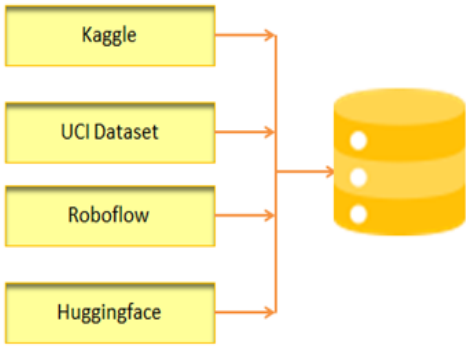


Fig. 4. Aggregation of raw dataset from multiple online sources.

The dataset used in this research consists of labelled mammographic images categorized into two clinically significant classes: Benign and Malignant, as illustrated in the figure. Benign images represent non-cancerous abnormalities, and malignant images correspond to confirmed cancerous cases. Fig. 3 describes the whole process of dataset collection. The mammograms exhibit variations in shape, texture, and tissue density, which are critical features for effective breast cancer detection. This diversity makes the dataset well-suited for training and evaluating machine learning models, as it enables them to learn discriminative patterns associated with different breast conditions. The dataset plays a crucial role in improving the robustness and accuracy of the proposed ML-based breast cancer detection system. The dataset used in this study was collected from two medical institutions in Bangladesh: Cumilla Moon Hospital and Cumilla Trauma Centre. A total of over 1,000 mammographic images were obtained with proper authorization and anonymization to ensure patient privacy. The collected data include cases classified as benign or malignant based on clinical diagnoses provided by experienced medical professionals. These real-world clinical images reflect a wide range of breast tissue characteristics, imaging conditions, and pathological variations, making the dataset representative of practical diagnostic scenarios (see Fig. 5).

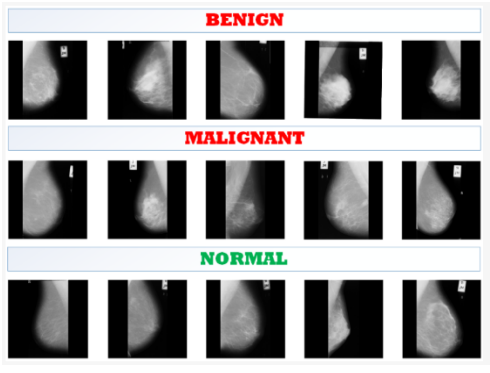


Fig. 5. Mammogram samples: Benign, malignant, and normal [4].

This dataset forms the foundation for training, validation, and performance evaluation of the proposed machine learning–based breast cancer detection models.

B. Image Preprocessing

Several essential preprocessing operations were performed internally within the convolutional neural network (CNN) models as part of the learning pipeline. The input images were first resized to match the fixed input dimensions required by the network architectures. Pixel intensity scaling and normalization were handled automatically within the model to stabilize gradient propagation and accelerate convergence during training. Convolutional layers learned hierarchical feature representations directly from the raw images, progressively capturing low-level features such as edges and textures, followed by higher-level semantic patterns.

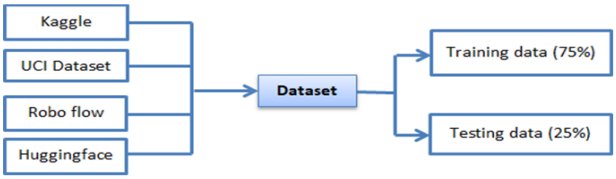


Fig. 6. Data preparation

Fig. 6 depicts the data preparation steps of this system, where, after data collection, a dataset was created. Then, the dataset splitting was done. The processed data fit the model. Here, data is gathered from different sources, which include ultrasound reports of the breast, and it is a very common report in breast cancer detection, where the patient's report is evaluated. Data is also collected from a medical-based website. This dataset was partitioned into two subsets for model development and evaluation. Specifically, 75% of the data was allocated for training, while the remaining 25% was reserved for testing. This data splitting strategy was employed to ensure effective model learning while enabling unbiased performance evaluation on unseen data.

Additionally, normalization layers embedded within the networks, such as batch normalization, mitigated internal covariate shift by standardizing intermediate feature distributions during training. Regularization mechanisms, including dropout layers, were employed to reduce overfitting and improve model generalization. Where applicable, data augmentation operations, such as random rotation, flipping, scaling, and intensity variation, were incorporated into the training pipeline and applied dynamically to input images. By integrating all preprocessing-related operations within the CNN framework, the proposed approach enables end-to-end learning, minimizes manual intervention, and enhances robustness across diverse imaging conditions.

C. Model Development

In this research, a convolutional neural network (CNN) is used as a baseline model to automatically classify breast images as benign or malignant, using both ultrasound and mammography datasets. CNNs are especially well-suited for medical images because they can learn hierarchical features directly from raw pixels, reducing the need for manual feature extraction. When a breast image is fed into the CNN, the network assigns learnable weights to small local regions, enabling it to capture important patterns such as mass boundaries, texture changes, and variations in tissue density. Compared to traditional machine learning methods, CNNs require less preprocessing and are highly effective at learning discriminative features from complex medical images.

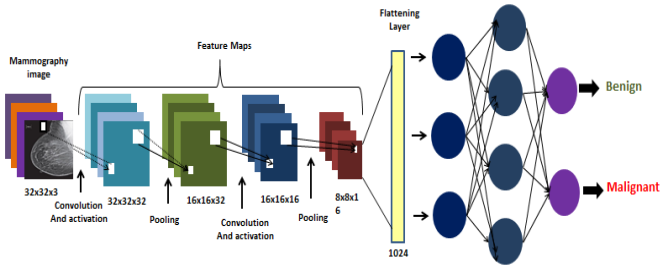


Fig. 7. Layer-wise architecture of the proposed CNN for mammography image classification. The network comprises three convolutional blocks with 32, 64, and 128 filters, followed by batch normalization, max-pooling, flattening, and fully connected layers with dropout regularization. The model was trained on a learning rate of 0.001 and 10 epochs. The final output layer performs binary classification into benign and malignant classes.

The CNN architecture used in this study follows a clear workflow, as shown in Fig. 7. It starts with an input layer that accepts preprocessed ultrasound or mammography images. These images then pass through several convolutional layers with learnable kernels, which automatically extract features ranging from low-level patterns like edges and contours to higher-level details such as lesion-specific textures. Nonlinear activation functions are applied after each convolution to give the network more flexibility in representing complex patterns. Pooling layers follow, reducing the size of the feature maps while keeping the important information, which helps lower computational costs and makes the model more robust to shifts or small changes in the images. The resulting feature maps are then flattened and passed through fully connected layers, which combine the learned features to make predictions. Finally, an output layer with a softmax activation produces probability scores for the benign and malignant classes.

On the ultrasound dataset, the baseline CNN achieved an accuracy of 85%, indicating that it can effectively capture basic texture and structural information. For mammography images, the CNN performed even better, reaching 97% accuracy, with an F1-score of 97%, macro-average of 96%, and weighted-average of 97%. These results highlight that a well-designed shallow CNN can handle high-quality mammograms, in which lesion boundaries and tissue density differences are more pronounced.

To further evaluate performance, AlexNet was also tested on these medical images. Thanks to its large convolutional kernels and deep fully connected layers, AlexNet is particularly good at capturing coarse texture features. On ultrasound images, AlexNet achieved the highest accuracy of 89%, outperforming both the baseline CNN and DenseNet. This suggests that AlexNet's design is well-suited for ultrasound data, where speckle noise and low-contrast patterns make feature extraction more challenging.

AlexNet is one of the early deep convolutional neural networks, with five convolutional layers and three fully connected layers. It is particularly good at capturing coarse texture features in medical images, especially ultrasound scans. Its large

kernels, ReLU activations, dropout layers, local response normalization, and data augmentation all work together to make learning more robust, even in the presence of speckle noise and low-contrast patterns. Designed to take full advantage of GPU processing, AlexNet outperformed baseline CNNs and DenseNet on ultrasound images, achieving 89% accuracy. This success comes from its deeper architecture, faster convergence, strong regularization, and ability to efficiently handle large datasets.

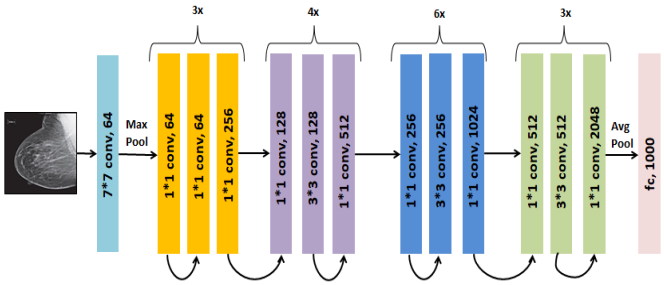


Fig. 8. Customized ResNet-50 architecture employed for mammographic image classification. The model consists of an initial 7×7 convolutional layer with 64 filters followed by max pooling, and four residual stages containing 3, 4, 6, and 3 bottleneck blocks, respectively. Each stage uses 1×1 and 3×3 convolutions with increasing feature dimensions (64–2048) to extract hierarchical features from mammograms. Global average pooling and a fully connected layer with 1000 neurons are used to produce the final classification output.

ResNet50 uses residual learning to make training deep networks more stable and efficient. On mammography images, it achieved an accuracy of 94%, with an F1-score of 93%, a macro-average of 93%, and a weighted-average of 94%. While its residual connections improved feature learning compared to VGG-based models, ResNet50 did not outperform the baseline CNN, suggesting that adding more depth isn't always necessary for mammogram classification.

The ResNet50 architecture, shown in Fig. 8, is structured into four main parts: convolutional layers, identity blocks, convolutional blocks, and fully connected layers. The convolutional layers first extract features from the input image. The identity and convolutional blocks then process and refine these features, while the fully connected layers make the final classification.

ResNet101 takes this a step further by increasing the network's depth to capture more complex patterns. It achieved slightly better results on mammography images, with an accuracy of 95% and F1 and average metrics around 95–96%. However, the performance gain over ResNet50 was marginal, suggesting that simply increasing depth can introduce redundancy without significantly improving diagnostic accuracy.

Fig. 9 shows the architecture of a customized ResNet-101–based CNN used for classifying mammography images. The process starts with a 7×7 convolutional layer with 64 filters, followed by max pooling to capture low-level features and reduce the spatial size of the image. The network then goes through four residual stages made up of identity blocks, which allow the deep architecture to train efficiently. The Conv2, Conv3, Conv4, and Conv5 stages contain 3, 4, 23, and 3 blocks, with channel depths increasing from 256 to

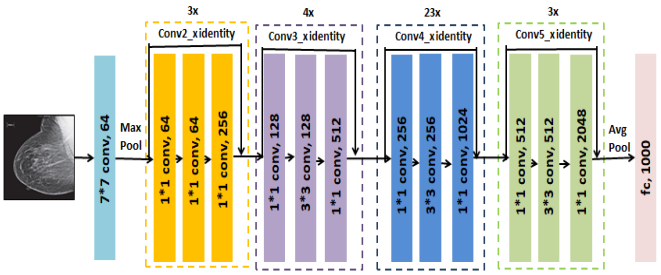


Fig. 9. Architecture of the customized ResNet-101–based CNN used for mammography image classification. The network begins with a 7×7 convolution and max-pooling, followed by four residual stages comprising Conv2 (3×), Conv3 (4×), Conv4 (23×), and Conv5 (3×) identity blocks with increasing channel depths of 256, 512, 1024, and 2048, respectively. Global average pooling and a fully connected layer with 1000 units are used for final classification.

2048. This structure helps the network progressively learn richer and more complex feature representations. After these stages, global average pooling condenses the information into a compact feature vector, which is then passed to a fully connected layer with 1000 units for classification.

MobileNetV3, on the other hand, is a lightweight model designed for efficiency. On mammography images, it achieved 92% accuracy, an F1-score of 91%, and balanced macro and weighted averages of 92%. While its performance is slightly lower than that of deeper residual models, MobileNetV3 offers a good balance between accuracy and computational cost, making it ideal for resource-constrained settings.

For experimental consistency, all deep learning models were trained using the Adam optimizer for 10 epochs with a learning rate of 0.001. The experiments revealed interesting trends across imaging types. Ultrasound images benefit from models that capture texture-rich features, such as AlexNet and DenseNet. Mammography images, with their clearer structures and higher contrast, consistently achieved better classification performance. Interestingly, the baseline CNN outperformed some of the deeper architectures on mammograms, suggesting that adding more depth doesn't always yield better results. Finding the optimal model complexity is key to balancing accuracy and efficiency.

D. Training and Evaluation

The training and evaluation process was designed to ensure robust learning, fair comparison among models, and reliable performance assessment across different imaging modalities. The collected dataset consists of ultrasound reports and mammography images, which were handled independently to preserve modality-specific characteristics. Each dataset was partitioned into training and testing subsets using a 75:25 split, ensuring sufficient data for model learning while retaining an unbiased test set for evaluation.

Prior to training, all images were resized to a uniform input dimension compatible with the selected network architectures. Pixel intensity values were normalized to improve convergence and stabilize the learning process. Labels were assigned according to clinical diagnosis, categorizing images into benign

and malignant classes. Data shuffling was applied to minimize ordering bias during training.

For the ultrasound dataset, deep learning models, including CNN, AlexNet, and DenseNet, were trained. In contrast, the mammography dataset was evaluated using various architectures, including CNN, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3. Transfer learning was employed where applicable by initializing networks with pre-trained weights and fine-tuning the higher layers to adapt to breast cancer imaging features. This strategy reduced training time and improved generalization, particularly for datasets with limited samples.

Model training was carried out using the Adam optimizer, with categorical cross-entropy as the loss function. The learning rate was empirically tuned to achieve stable convergence, and early stopping was applied to prevent overfitting by monitoring validation loss. Batch size and number of epochs were selected based on computational efficiency and model performance during preliminary experiments:

$$P = \frac{TP}{TP + FP} \tag{1}$$

To evaluate classification effectiveness, multiple performance metrics were employed, including accuracy, precision, recall, and F1-score, which collectively provide a balanced assessment of model reliability, especially in the presence of class imbalance. Accuracy measures overall correctness, while precision and recall quantify false-positive and false-negative behavior, respectively. where TP and FP represent true positives and false positives, respectively. The recall R evaluates the ratio of correctly detected damaged regions to the total number of actual damaged regions:

$$R = \frac{TP}{TP + FN} \tag{2}$$

where, FN denotes false negatives.

The F1-score, which is the harmonic mean of precision and Recall, provided a balanced assessment of detection:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{3}$$

The F1-score is a harmonic mean of precision and recall, providing a comprehensive indicator of diagnostic performance. Comparative evaluation was conducted across all trained models to identify the most effective architecture for each imaging modality. The final prediction stage classifies input images as either benign or malignant, thereby supporting clinical decision-making with consistent and interpretable results. The evaluation framework ensures that the proposed system is both robust and generalizable, demonstrating its suitability for real-world breast cancer diagnosis applications.

IV. Results and Discussion

This section analyzes the performance of the proposed models using both quantitative evaluation metrics and qualitative visual inspection. Quantitative results highlight numerical performance differences, while qualitative analysis, as shown in confusion matrices, provides insight into classification consistency and misclassification patterns.

A. Result Analysis

1) Quantitative performance analysis: The quantitative evaluation of the framework is presented through performance metrics obtained from mammography and ultrasound datasets, as summarized in Table I and Table II, respectively. The models were assessed using accuracy, precision, recall, and F1-score to ensure a comprehensive evaluation under class imbalance conditions.

TABLE I. Performance Comparison of Deep Learning Models on the Mammography Dataset (Precision, Recall, F1-Score are Written as Benign/Malignant Format).

Model	Precision	Recall	F1-Score	Accuracy
CNN	97% / 96%	95% / 97%	96% / 97%	97%
ResNet50	92% / 95%	94% / 93%	93% / 94%	94%
ResNet101	95% / 95%	94% / 96%	95% / 96%	95%
VGG16	91% / 90%	87% / 93%	89% / 92%	90%
VGG19	91% / 90%	87% / 93%	89% / 91%	90%
MobileNetV3	88% / 95%	94% / 90%	91% / 92%	92%

For the mammography dataset (Table I), the CNN model achieved the highest overall performance with an accuracy of 97%, demonstrating balanced precision and recall for both benign and malignant classes (precision: 97%/96%, recall: 95%/97%). This indicates strong discriminatory capability and stable generalization. ResNet101 performed closely, achieving 95% accuracy of 95%, benefiting from residual learning to capture complex mammographic patterns. ResNet50 achieved a slightly lower accuracy of 94%, suggesting that deeper residual connections in ResNet101 contribute to improved feature representation.

VGG-based architectures (VGG16 and VGG19) exhibited comparatively lower accuracy (90%), primarily due to reduced recall for benign cases (87%), indicating higher false-negative rates. MobileNetV3 achieved an accuracy of 92%, offering a favorable trade-off between performance and computational efficiency, particularly with high malignant recall (94%), which is clinically significant for cancer detection.

TABLE II. Accuracy Comparison of Models on the Ultrasound Dataset.

Model	Accuracy
CNN	85%
AlexNet	89%
DenseNet	88%

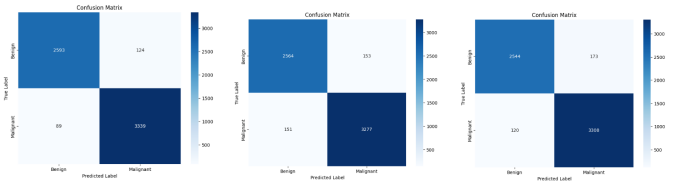
For the ultrasound dataset (Table II), overall classification performance was lower compared to mammography, reflecting the inherent challenges of ultrasound imaging, such as noise, speckle artifacts, and low contrast. Among the evaluated models, AlexNet achieved the highest accuracy of 89%, followed by DenseNet at 88% and CNN at 85%. The improved performance of AlexNet and DenseNet indicates their effectiveness in capturing texture-based features from ultrasound images.

Overall, the quantitative results confirm that mammography-based classification outperforms ultrasound-based classification across all evaluated models. The superior

performance of CNN and ResNet101 on mammographic images highlight the importance of deep hierarchical feature learning for high-resolution medical imaging. The performance reporting in this study is not fully symmetric across modalities. For the mammography dataset, detailed evaluation metrics including precision, recall, F1-score, and accuracy are reported, whereas for the ultrasound dataset, only accuracy values are presented. This difference arises from the original experimental setup and the available evaluation outputs. Although this limits a strictly uniform cross-modal comparison, the results still provide meaningful insights into model performance trends across different imaging modalities. Future work will focus on extending the evaluation to include comprehensive metrics for all datasets to enable a fully balanced comparison. These findings demonstrate the robustness and clinical relevance of the proposed deep learning framework for automated breast cancer diagnosis.

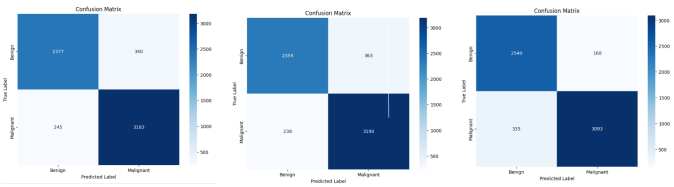
2) Qualitative performance analysis: Qualitative analysis is performed to visually evaluate the classification behavior of the proposed deep learning models beyond numerical performance metrics. Confusion matrices provide intuitive insight into how well each model distinguishes between benign and malignant mammographic images, highlighting both correct predictions and patterns of misclassification.

As illustrated in Fig. 10 and Fig. 11, the confusion matrices of all six models exhibit strong diagonal dominance, indicating a high number of correctly classified samples for both benign and malignant classes. The CNN-based model shows a relatively balanced distribution of true positives and true negatives, with only a limited number of false classifications. This suggests that the model effectively captures essential features required for mammogram classification.



(a) CNN (b) ResNet50 (c) ResNet101

Fig. 10. Confusion matrix for CNN, ResNet50 and ResNet 101 on the mammography image.



(a) VGG16 (b) VGG19 (c) MobileNetv3

Fig. 11. Confusion matrix for VGG16, VGG19 and MobileNetv3 on the mammography image.

The MobileNetV3 confusion matrix demonstrates improved detection of malignant cases, as reflected by a higher count of true positives, although a small number of benign cases are miResNet-based architectures, these models exhibit a slightly increased misclassification of benign cases, which may be attributed to their architectural complexity and the achieves the most consistent separation between the two classes, highlighting the advantage of deeper residual learning in minimizing classification errors.

The VGG16 and VGG19 models also present strong performance, with a high proportion of malignant samples correctly identified. However, compared to ResNet-based architectures, these models exhibit slightly higher misclassification rates of benign cases, which may be attributed to their greater architectural complexity and the absence of residual connections.

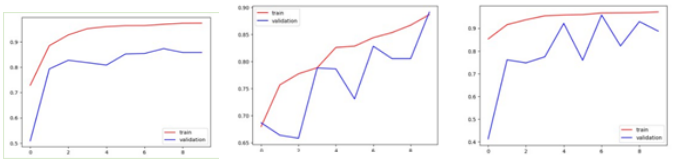
Overall, the visual analysis confirms that all evaluated models are capable of effective mammogram classification, with ResNet101 and CNN providing the most balanced and reliable predictions. The qualitative observations are consistent with the quantitative results, reinforcing the effectiveness of deep learning approaches for breast cancer detection.

B. Discussion

This study presents a comparative analysis of multiple CNN-based architectures for breast cancer classification, utilizing both ultrasound and mammography imaging modalities. A clear performance gap is observed between the two imaging modalities. Overall, mammography-based models consistently outperform ultrasound-based models, achieving higher accuracy, F1-score, macro-average, and weighted-average values. Although the proposed framework achieves strong classification performance, the current study does not evaluate computational complexity, training overhead, inference latency, memory consumption, or deployment cost. These factors are important for real-world clinical deployment, where resource efficiency and scalability are critical. Therefore, a comprehensive evaluation of these aspects is considered as a potential direction for future work. This outcome can be attributed to the higher contrast, clearer structural details, and more consistent appearance of lesions in mammography images compared to ultrasound images, which often suffer from speckle noise and low contrast.

1) Ultrasound-based model performance: For the ultrasound dataset, AlexNet achieved the highest accuracy of 89%, followed closely by DenseNet at 88%, while the baseline CNN achieved 85% accuracy. As shown in Fig. 12(b) and Fig. 12(c), these models demonstrate faster convergence during training but exhibit noticeable fluctuations in validation accuracy. Such instability reflects the inherent variability and noise present in ultrasound images.

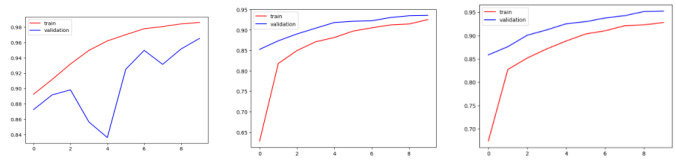
AlexNet's superior performance suggests that architectures with larger convolutional kernels and strong texture-extraction capabilities are more effective for ultrasound imaging. DenseNet also performed competitively due to its dense feature reuse mechanism; however, the increased architectural complexity did not translate into significant performance gains. The baseline CNN, while simpler, showed limited capability in modeling the complex texture patterns characteristic of ultrasound images [see Fig. 12(a)].



(a) CNN (b) AlexNet (c) DenseNet

Fig. 12. Training and validation accuracy curves for CNN, AlexNet, and DenseNet on the ultrasound dataset.

2) Mammography-based model performance: In contrast, mammography-based experiments yielded substantially higher and more stable performance across all models. Notably, the baseline CNN achieved the best overall performance, with an accuracy of 97%, an F1-score of 97%, a macro-average of 96%, and a weighted-average of 97%. As illustrated in Fig. 13(a), the CNN model exhibits smooth convergence and minimal divergence between training and validation curves, indicating strong generalization and low overfitting.



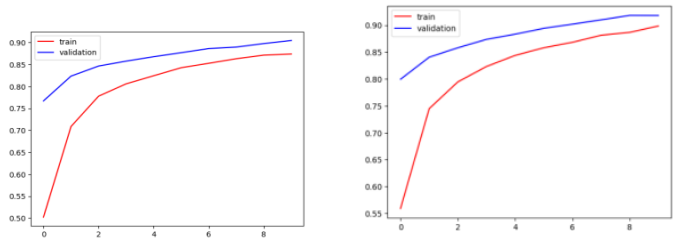
(a) CNN (b) ResNet50 (c) ResNet101

Fig. 13. Training and validation accuracy curves for CNN, ResNet50, and ResNet101 on the mammography dataset.

Among deeper architectures, ResNet101 achieved 95% accuracy, marginally outperforming ResNet50 (94%). Fig. 13(b) and Fig. 13(c) show that residual networks benefit from stable training dynamics due to skip connections, yet the additional depth of ResNet101 provides only limited improvement over ResNet50. This suggests diminishing returns from increased network depth when applied to high-quality mammographic data.

3) Performance of VGG and mobilenet architectures: VGG16 and VGG19 achieved moderate performance with accuracies of 90%, as shown in Fig. 14(a) and corresponding results in the table. Although VGG architectures are effective feature extractors, their large parameter count and lack of residual connections may contribute to overfitting and reduced generalization in medical imaging tasks.

MobileNetV3 achieved an accuracy of 92%, offering a favorable balance between performance and computational efficiency. As shown in Fig. 14(b), MobileNetV3 demonstrates stable learning behavior, making it a practical choice for deployment in resource-constrained clinical environments, despite slightly lower accuracy compared to deeper residual models. These findings emphasize that architecture selection should be guided by data characteristics rather than depth alone, particularly in medical imaging applications. The strong performance and stable generalization of the baseline CNN on mammography images highlight its potential for integration into computer-aided diagnosis (CAD) systems. Such systems



(a) VGG16 (b) MobileNetv3

Fig. 14. Training and validation accuracy curves for VGG16 and MobileNetv3 on the mammography dataset.

can assist radiologists by providing reliable second opinions, improving early detection rates, and ultimately enhancing patient outcomes.

4) Comparative analysis with existing studies: Previous papers on breast cancer detection have predominantly relied on traditional machine learning techniques, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forests, and Logistic Regression, often using handcrafted features and benchmark datasets, including the Wisconsin Breast Cancer Dataset or limited histopathological image collections. While these approaches have demonstrated reasonable classification accuracy, their effectiveness is constrained by manual feature engineering, limited scalability to large image datasets, and reduced robustness when handling complex visual patterns and class imbalance.

In contrast, the proposed study adopts a deep learning–based approach that eliminates the need for handcrafted feature extraction by enabling automatic hierarchical feature learning through convolutional neural networks. Unlike earlier works that focus on a single imaging modality, this research performs a comparative analysis of both ultrasound and mammography images, providing a more comprehensive evaluation of modality-specific diagnostic performance.

Moreover, several prior studies report results based on a single or limited number of classifiers, whereas the proposed framework evaluates multiple CNN architectures, including CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3. This comprehensive model comparison enables a deeper understanding of how network depth, residual learning, and lightweight architectures impact breast cancer classification accuracy.

Another key limitation of earlier research is the reliance on public or small-scale datasets, which may not accurately reflect real-world clinical scenarios. The proposed study addresses this gap by validating models using real-world ultrasound and mammography datasets, thereby enhancing the clinical relevance and generalizability of the findings.

Performance-wise, the proposed framework achieves superior accuracy and F1-score, particularly in mammography-based classification, where the CNN and ResNet101 models outperform traditional machine learning methods reported in prior literature. Additionally, the use of macro and weighted-average metrics ensures a more reliable evaluation under conditions of class imbalance, a consideration often overlooked

in earlier studies.

Overall, the proposed research advances existing work by offering a modality-aware, deep learning–driven, and clinically relevant framework that demonstrates improved diagnostic performance and robustness, making it more suitable for practical deployment in computer-aided breast cancer diagnosis systems.

V. Conclusion

This study has presented a deep learning–based framework for automated breast cancer classification using ultrasound and mammography images. Multiple CNN architectures, including CNN, AlexNet, DenseNet, ResNet50, ResNet101, VGG16, VGG19, and MobileNetV3, were systematically evaluated to analyze modality-specific performance. The experimental results demonstrate that deep learning models effectively learn discriminative features directly from medical images, eliminating the need for handcrafted feature extraction and outperforming traditional machine learning approaches.

Mammography-based classification achieved superior performance compared to ultrasound imaging, with CNN and ResNet101 producing the highest accuracy and F1-scores, highlighting the effectiveness of deeper architectures in capturing complex structural patterns. The present study does not include robustness evaluation under challenging conditions such as noisy inputs, image corruption, or domain shift scenarios. As a result, the stability and generalization ability of the proposed models under real-world variations in medical imaging data remain unverified. Future research will focus on addressing these issues by incorporating robustness testing and domain adaptation techniques. Although ultrasound-based results were comparatively lower due to inherent image noise and low contrast, AlexNet and DenseNet exhibited stable learning behavior and competitive performance. Overall, the proposed framework offers a robust and clinically relevant computer-aided diagnostic solution that can aid radiologists in early breast cancer detection and informed decision-making.

Future research will focus on expanding the framework using larger multi-institutional datasets to improve generalization. Incorporating advanced data augmentation techniques, class imbalance handling methods, attention mechanisms, and tumor localization or segmentation modules may further enhance diagnostic accuracy and interpretability. Additionally, real-time deployment and clinical validation remain crucial directions for the practical adoption in healthcare settings.

References

[1] "Breast cancer," LCH, https://labaidcancer.com/cancer-details/Breast-Cancer (accessed Dec. 19, 2025). .

[2] H. Australia, "Breast cancer," Healthdirect Australia. Accessed: Dec. 20, 2025. [Online]. Available: https://www.healthdirect.gov.au/breast-cancer

[3] J. Ferdause, A. K. M. S. Kadir, and M. Haque, "Understanding diagnostic delays among newly diagnosed breast cancer patients at a tertiary cancer care center in a low-middle-income country like Bangladesh," *Medicine*, vol. 104, e41775, 2025, doi: 10.1097/MD.0000000000041775.

[4] S. Charan, J. Khan, and K. Khurshid, "Breast cancer detection in mammograms using convolutional neural network," in Proc. IEEE Int. Conf. Mathematics and Engineering Technologies (iCoMET), Mar. 2018, doi: 10.1109/ICOMET.2018.8346384.

[5] C. Easttom, S. Thapa, and J. Lawson, "A comparative study of machine learning algorithms for use in breast cancer studies," in Proc. 2020 IEEE Annu. Comput. Commun. Workshop Conf. (CCWC), Jan. 2020, pp. 412–416, doi: 10.1109/CCWC47524.2020.9031266.

[6] M. Ranjan, A. Shukla, K. Soni, S. Varma, and M. Kuliha, "Cancer prediction using random forest and deep learning techniques," in Proc. 2022 Int. Conf. Communication Systems and Network Technologies (CSNT), 2022, pp. 227–231, doi: 10.1109/CSNT54456.2022.9787608.

[7] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in Proc. 2018 Int. Conf. on Trends in Electronics and Informatics (CTEMS), 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.

[8] T. Jain, V. K. Verma, A. Yadav, and A. Jain, "Supervised machine learning approach for the prediction of breast cancer," in Proc. IEEE Int. Conf. System, Computation, Automation and Networking (ICSCAN), MVIT Puducherry, India, Jul. 3–4, 2020, doi: 10.1109/ICSCAN49426.2020.9262403.

[9] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in Proc. 2016 5th Int. Conf. Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, United Arab Emirates, Dec. 6–8, 2016, pp. 1–4, doi: 10.1109/ICEDSA.2016.7818560.

[10] A. Joshi and A. Mehta, "Mammogram image classification using various machine learning algorithms," in Proc. 2022 7th Int. Conf. on Computing, Communication and Security (ICCCS), Nov. 2022, pp. 106–110, doi: 10.1109/ICCCS55188.2022.10079398.

[11] B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," in Proc. 2016 IEEE Int. Conf. Computational Intelligence and Computing Research (ICCIC), Chennai, India, Dec. 2016, pp. 1–5, doi: 10.1109/ICCIC.2016.7919576.

[12] N. Khuriwal and N. Mishra, "Breast cancer detection from histopathological images using deep learning," in Proc. 2018 3rd Int. Conf. and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Nov. 2018, pp. 1–4, doi: 10.1109/ICRAIE.2018.8710426.

[13] A. Kumar, R. Patra, and A. Ghosh, "Model selection for predicting breast cancer using supervised machine learning algorithms," in Proc. 2020 IEEE 1st Int. Conf. for Convergence in Engineering (ICCE), Kolkata, India, 2020, pp. 320–324, doi: 10.1109/ICCE50343.2020.9290578.

[14] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning," in Proc. 2018 37th Chinese Control Conference (CCC), July 2018, pp. 9428–9433, doi: 10.23919/ChiCC.2018.8483140.

[15] E. Halim, P. P. Halim, and M. Hebrard, "Artificial intelligent models for breast cancer early detection," in Proc. 2018 Int. Conf. Information Management and Technology (ICIMTech), Indonesia, 2018, pp. 517–521, doi: 10.1109/ICIMTech.2018.8528140.

[16] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in Proc. 2018 Int. Conf. on Computational Techniques in Engineering and Management (CTEMS), 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.

[17] M. R. Ahmed, M. A. Ali, J. Roy, S. Ahmed, and N. Ahmed, "Breast cancer risk prediction based on six machine learning algorithms," in Proc. 2020 IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, Dec. 2020, pp. 1–5, doi: 10.1109/CSDE50874.2020.9411572.

[18] V. A. Telsang and K. Hegde, "Breast cancer prediction analysis using machine learning algorithms," in Proc. 2020 Int. Conf. on Communication, Computing and Industry 4.0 (C2I4), Dec. 2020, pp. 1–5, doi: 10.1109/C2I451079.2020.9368911.

[19] S. Bhise, S. Gadekar, A. S. Gaur, S. Aswale, et al., "Detection of breast cancer using machine learning and deep learning methods," in Proc. 2022 3rd Int. Conf. Intelligent Engineering and Management (ICIEM), Apr. 2022, doi: 10.1109/ICIEM54221.2022.9853080.

[20] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast cancer detection using K-nearest neighbor machine learning algorithm," in Proc. 2016 9th Int. Conf. on Developments in eSystems Engineering (DeSE), Liverpool, United Kingdom, Aug.–Sep. 2016, pp. 35–39, doi: 10.1109/DeSE.2016.8.

[21] T. Mehejabin, F. Rahman, S. Yeasmin, M. Sarkar, and F. Rahman, "Identification of most relevant breast cancer miRNA using machine learning algorithms," in Proc. 2020 IEEE 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–6, doi: 10.1109/ICCCNT49239.2020.9225624.

[22] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in Data Mining and Big Data, Y. Tan, Y. Shi, and M. Tuba, Eds., Communications in Computer and Information Science, vol. 1234, Springer, Singapore, Jul. 2020, pp. 108–117, doi: 10.1007/978-981-15-7205-0-10.

[23] F. Rahman, T. Mehejabin, S. Yeasmin, and M. Sarkar, "A comprehensive study of machine learning approach on cytological data for early breast cancer detection," in Proc. 2020 IEEE 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–6, doi: 10.1109/ICCCNT49239.2020.9225448.

[24] D. Selvathi and A. Poornila, "Performance analysis of various classifiers on deep learning network for breast cancer detection," in Proc. 2017 Int. Conf. on Signal Processing and Communication (ICSPC), Coimbatore, India, Jan. 2017, pp. 359–363, doi: 10.1109/CSPC.2017.8305869.

[25] A. R. Vaka, B. Soni, and S. Reddy K., "Breast cancer detection by leveraging machine learning," ICT Express, vol. 6, no. 4, pp. 320–324, Dec. 2020, doi: 10.1016/j.icte.2020.04.009.

[26] X. Zhou, Y. Li, R. Gururajan, G. Bargshady, X. Tao, R. Venkataraman, P. D. Barua, and S. Kondalsamy-Chennakesavan, "A new deep convolutional neural network model for automated breast cancer detection," in Proc. 2020 7th IEEE Int. Conf. on Behavioural and Social Computing (BESC), Bournemouth, United Kingdom, 5–7 Nov. 2020, pp. 1–4, doi: 10.1109/BESC51023.2020.9348322.

[27] L. Qian, J. Bai, Y. Huang, D. Q. Zeebaree, A. Saffari, and D. A. Zebari, "Breast cancer diagnosis using evolving deep convolutional neural network based on hybrid extreme learning machine technique and improved chimp optimization algorithm," Biomedical Signal Processing and Control, vol. 87, p. 105492, Jan. 2024, doi: 10.1016/j.bspc.2023.105492.

Optimization of Service Function Chain Placement in Cloud-Fog-Edge Networks

Chandrapal Singh Dangi, Sanjay Sharma
Department of Mathematics, Bioinformatics and Computer Applications,
Maulana Azad National Institute of Technology, Bhopal, India

***Abstract*—There is an explosion in IoT devices, 5G technology, and MECs, that results in increasing demands on effective and scalable network services management. Service function chaining, defined as the sequence of functions in VNFs on a path, is one of the core principles behind the NFV architecture design. SFC allocation to the heterogeneous clouds–fogs–edges network is an NP-hard problem characterized by mutually conflicting goals, such as latency minimization, energy and cost reduction, and resource maximization. In this study, an in-depth comparison study is carried out on three population-based optimization algorithms for solving the placement problem of SFCs using Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Grey Wolf Optimization (GWO) on three cases: (1) VNF deployment cost/QoE optimization in a 5G hybrid cloud with 12 nodes and weighting factor γ=0.4; (2) SFC graph matching on MEC-NFV networks with a 100-node physical network, 20 VNFs, and equal utilization weights α=β=γ=1/3; and (3) multi-instance SFC mapping on Fog-to-Cloud (F2C) IoT environment with a 5-VNF chain across 5 nodes. These three algorithms have been evaluated under identical conditions: 10 independent runs, 200 iterations, 20–30 agents. Results demonstrate that GWO achieves the best VNF deployment objective (W=73.92, a 15.1% improvement over the BGWO baseline), PSO achieves the highest resource utilization (52.9%) in MEC-NFV placement, and both PSO and GWO reduce F2C end-to-end latency by 25% compared to the ILP reference (12 vs. 16 units at three instances), while all three algorithms reduce latency by approximately 80% relative to cloud-only deployment. PSO emerges as the most consistently high-performing algorithm across all three scenarios.**

***Keywords*—Service function chaining; Virtual Network Functions; Network Function Virtualization; Particle Swarm Optimization; Ant Colony Optimization; Grey Wolf Optimization; mobile edge computing; Fog-to-Cloud; 5g networks; resource allocation**

I. Introduction

The convergence of cloud computing, edge intelligence, and programmable networking has fundamentally transformed how network services are designed, deployed, and managed. Modern communication infrastructures increasingly rely on Network Function Virtualization (NFV) to decouple network functions traditionally implemented in dedicated hardware from their physical substrates. This paradigm enables deploying network services as software modules, known as Virtual Network Functions (VNFs), hosted on commodity servers in cloud data centers, edge nodes, or fog computing layers [1], [2].

A *Service Function Chain* (SFC) is an ordered sequence of VNFs through which network traffic must traverse to receive a composite service. A typical SFC for online gaming, for instance, consists of a Network Address Translator (NAT), Firewall (FW), Video Optimization Controller (VOC), WAN Optimization Controller (WOC), and Intrusion Detection and Prevention System (IDPS), each applied sequentially to user traffic [3]. The problem of SFC allocation, which involves placing each VNF on its corresponding physical node and routing data to the respective node accordingly, forms the core issue addressed in this [4] research study.

The development of telecommunication technology, for instance, the fifth generation (5G) of mobile telecommunication networks, and new areas of research and application, such as the Internet of Things (IoT), alter the resource access paradigm of the CSPs. In order to cut down the access time and enhance the QoS, resources are placed near the cloud users, and therefore, the paradigm of mobile edge computing (MEC) has been introduced. The cloud users access the network resources via their Mobile Devices using the base station and obtain their service.

There are three main reasons why this problem is difficult. The first is that there are two competing goals, namely reducing latency which favors edge/fog computing and saving costs which favors the central cloud. The second reason is due to heterogeneity in the resources available at the nodes leading to complicated interactions of constraints. Lastly, the problem is NP-Hard [5].

This study seeks to examine whether PSO, ACO, and GWO algorithms can be used to allocate SFCs in three selected environments; 5G cloud environment [6], MEC-NFV [7], and F2C IoT. The 5G Cloud SFC focuses on high throughput and centralized management. The MEC-NFV SFC focuses on ultra-low latency and edge computing capabilities.The F2C IoT SFC focuses on distributed service placement, energy efficiency, and heterogeneity of IoT resources. Thus, SFC management is getting increasingly more decentralized and complicated when transitioning from 5G Cloud to MEC-NFV and further to F2C IoT.

Principal contributions:

- We propose a metaheuristic algorithm-based approach to solve 3 SFC architectures 5G Cloud Environment,MEC-NFV Environment and F2C (Fog-to-Cloud) IoT Environment.
- We have simulated all 3 SFC scenario in python programming version 3.11 and the simulation results demonstrated that GWO outperforms BGWO by +15.1%; PSO outperforms with 52.9% MEC-NFV efficiency; PSO/GWO cut down F2C delay by 25% relative to ILP.
- Consistent execution and systematic comparison of PSO, ACO, and GWO algorithms in three different

SFC configurations. Thorough testing under completely controlled conditions for valid statistical comparisons.

- Concrete algorithm selection recommendation for cloud-fog/edge systems.

II. Related Work

SFC placement and VNF deployment have attracted substantial research across multiple network paradigms. In MEC contexts, strategies range from QoS-aware placement algorithms targeting service availability and access latency [8] to dynamic schedulers that minimize VNF migrations as user mobility patterns evolve. Integer Linear Programming (ILP) and Mixed ILP formulations provide optimal solutions for small-scale SFC placement instances, but become computationally intractable as network size grows [9]. Column-generation approaches extend ILP tractability by decomposing the problem into configuration enumeration and selection phases, though these remain limited by quadratic constraint structures [10].

Metaheuristic algorithms have emerged as a dominant paradigm for large-scale SFC optimization. PSO has been applied to mobility-aware VNF placement enabling dynamic reallocation as users move between base stations. ACO has been used to balance communication delay against VNF relocation count through pheromone-trail encoding. Genetic Algorithms have demonstrated effectiveness in multi-objective formulations approximating latency-cost Pareto fronts [11]. Grey Wolf Optimization (GWO), inspired by the social hierarchy and hunting behavior of grey wolves, has been applied to VNF deployment in hybrid cloud infrastructures. The Binary GWO (BGWO) variant addresses the discrete nature of VNF assignment through sigmoid-based position discretization [12]. Comparative studies suggest GWO achieves faster convergence than genetic algorithms on similar combinatorial problems due to its adaptive exploration–exploitation balance.

In Fog-to-Cloud (F2C) computing, SFC mapping presents additional complexity from heterogeneous fog-to-cloud latency characteristics. Multi-instance SFC deployment has been shown to substantially reduce end-to-end latency by enabling traffic assignment to the nearest available instance [13]. Graph-matching formulations have been proposed for MEC-NFV placement, where similarity between the VNF forwarding graph and the physical network graph is maximized using LP followed by Hungarian algorithm-based mapping, reducing execution time and improving node utilization [14].

Recent advances in Service Function Chain (SFC) placement and Virtual Network Function (VNF) deployment have increasingly leveraged Deep Reinforcement Learning (DRL) and hybrid optimization techniques to address the challenges of dynamic and large-scale network environments. Zhou et al. [15] proposed a DRL-based multi-objective SFC placement framework that jointly optimizes resource utilization, latency, and service acceptance rates in NFV-enabled networks. Their approach demonstrated the effectiveness of reinforcement learning in adapting placement decisions to changing network conditions while improving overall service performance.

Feng et al. [16] introduced a hybrid Deep Reinforcement Learning-Greedy (DRL-G) algorithm for SFC deployment in space–air–ground integrated networks. The proposed framework combines the adaptive learning capability of DRL with the computational efficiency of greedy search to optimize deployment cost, energy consumption, and service latency. Experimental results showed improved scalability and convergence compared to conventional optimization methods.

In Li et al. [17], the authors suggested an optimized approach to cache and dynamically migrate services in the MEC environment by implementing a Deep Q-Network (DQN). They pointed out that the users' movement and insufficient MEC coverage result in disruption of services, higher latency, and increased energy costs. For solving these problems, they combined edge caching with intelligent migration of services.

To address distributed orchestration challenges in edge environments, Li et al. [18] developed a multi-agent reinforcement learning (MARL) framework for SFC placement in edge computing networks. By enabling multiple agents to collaboratively perform VNF deployment and path selection decisions, the proposed approach achieved higher resource utilization and service provider profitability while supporting decentralized network management.

In general, the literature presents a clear evolution in terms of moving from cloud placement to dynamic cloud-fog-edge placement of services that are QoS aware, energy efficient, and mobility aware. The surveys present very good taxonomies, but recent research on algorithms presents heuristic, approximations, game theory, and meta-heuristics approaches. Nevertheless, there are some issues that are not well covered, such as the combination of VNF placement with routing, dynamic multi-service function chain requests, user mobility, failure tolerance, scaling, and unified optimization of latency, energy, resource efficiency, throughput, and reliability.

Despite the breadth of existing work, systematic multi-scenario comparative studies evaluating the same set of metaheuristic algorithms across diverse environments under rigorously identical conditions remain scarce. This study addresses that gap.

III. System Model

A. Network Architecture

The study considers three representative deployment environments unified by the Fog-to-Cloud continuum: IoT devices connect to nearby Fog/Edge nodes, which link to regional edge clouds and a central cloud data center. SFC placement determines which layer hosts each VNF, directly controlling service latency, energy consumption, and resource utilization. Fig. 1 illustrates the SFC architecture and VNF chain structure.

B. Scenario I: 5G Hybrid Cloud

The first scenario models a 5G network with a hybrid cloud infrastructure comprising one central cloud server and eleven edge cloud servers (12 nodes total). Each edge cloud serves a set of evolved Node-B (eNB) base stations, with each eNB aggregating VNF requests from hundreds to thousands of mobile users. Inter-cloud latencies are 10–200 ms; eNB-to-edge latencies are 2–5 ms. Each edge cloud offers VMs with 500–1000 MIPS and 512–1024 MB RAM. VNF holding capacity ranges from 100–300 instances per node (600 for the central cloud). The weight factor γ=0.4 governs the trade-off between deployment cost and service delay.

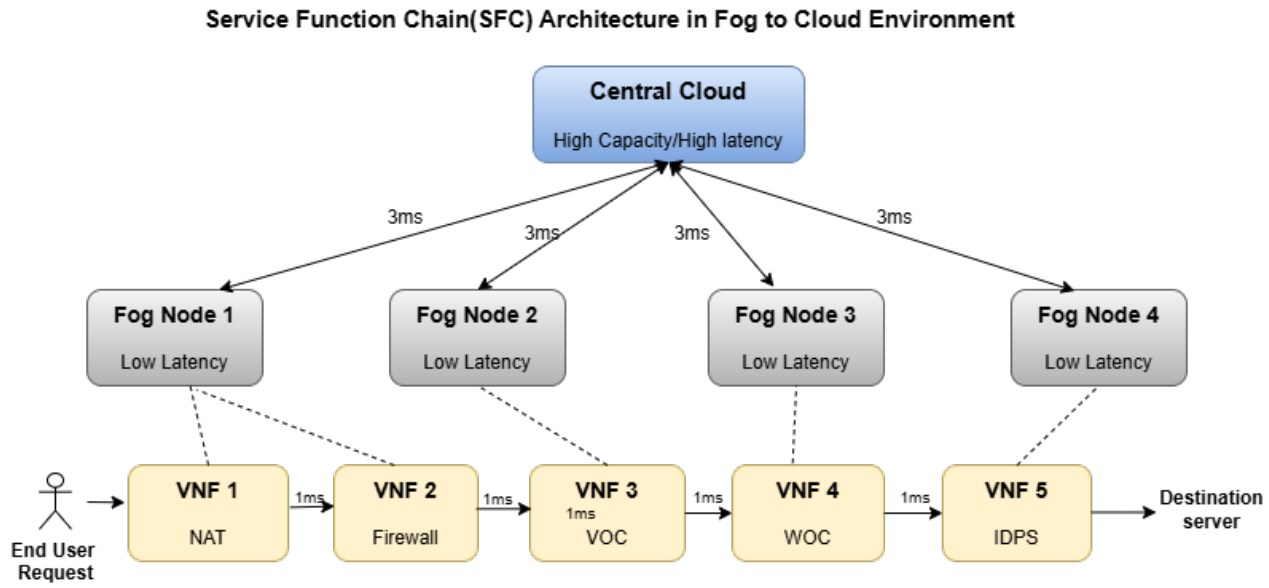


Fig. 1. SFC Architecture. IoT user traffic traverses an ordered chain of five VNFs (NAT, FW, VOC, WOC, IDPS) hosted across fog and cloud nodes. VNF-to-node assignment determines end-to-end latency.

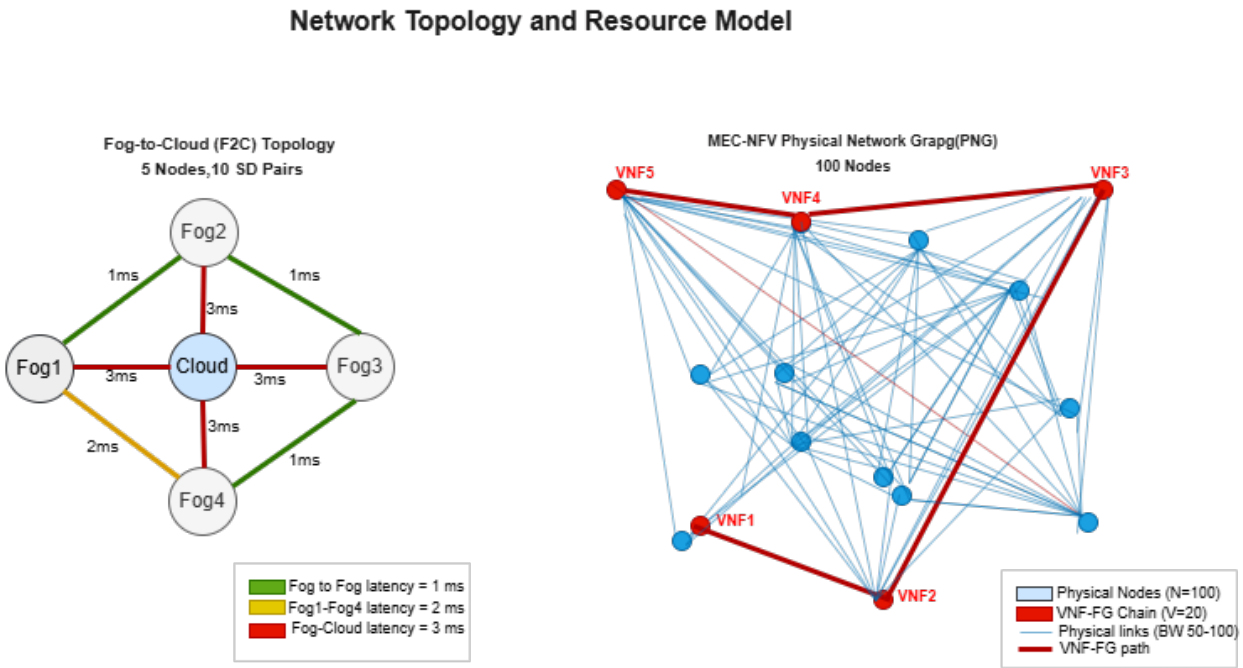


Fig. 2. Network topology models. *Left*: F2C topology (5 nodes). *Right*: Representative subgraph of the 100-node MEC-NFV physical network with the 20-VNF chain highlighted.

C. Scenario II: MEC-NFV Environment

The second scenario considers a three-layer MEC-NFV architecture. The physical network graph (PNG) has N=100 nodes with CPU and memory capacities uniformly in [50, 100] units, 50% connectivity (θ=0.5), and link bandwidths of 50–100 units. VNFs consume 1–20 units of CPU, memory, and bandwidth. The placement objective maximizes a weighted resource utilization with equal weights α=β=γ=1/3.

D. Scenario III: Fog-to-Cloud IoT

The third scenario implements the F2C architecture for IoT services. The topology consists of five nodes: four fog nodes in a mesh and one central cloud node (see Fig. 2). Link latencies follow: fog–fog = 1, fog1–fog4 = 2, fog–cloud = 3 (latency units). Ten source-destination pairs each carry 1 Gbps of traffic through a 5-function service chain (NAT→FW→VOC→WOC→IDPS). The system supports up to seven concurrent SFC instances.

E. Latency and Cost Models

The end-to-end service latency for VNF f assigned to cloud node k from eNB j is:

$$D(f,k,j) = R(f,k,j) + \Upsilon(f,k,j) + \phi(f,k,j) \tag{1}$$

where, $R(f,k,j)$ is the relocation (migration) delay, $\Upsilon(f,k,j)$ is the communication latency (eNB-to-cloud and inter-cloud), and $\phi(f,k,j) = I_j^f/e_k$ is the processing delay (instructions / MIPS).

The deployment cost $B(f,k,j)$ accounts for local vs. remote cloud pricing:

$$B(f,k,j) = \Gamma_{k,j}^f \cdot \vartheta_f \cdot \left[x_k^f \cdot l_k + (1 - x_k^f) \cdot r_k\right] \tag{2}$$

where, l_k and r_k are local and remote unit costs, and $x_k^f \in \{0,1\}$ indicates whether the VNF was previously hosted at node k.

IV. Problem Formulation

A. Scenario I: Multi-Objective VNF Deployment

The Min-CD problem minimizes a weighted combination of normalized deployment cost and normalized service delay:

$$\min\ W = \sum_{j\in M}\sum_{f\in\mathcal{F}_j}\sum_{k\in C}\Big\{\gamma\cdot B'(f,k,j)+(1-\gamma)\cdot D'(f,k,j)\Big\} \tag{3}$$

where, $B'(f,k,j) = B(f,k,j)/U_j^f$ and $D'(f,k,j) = D(f,k,j)/\Delta_j^f$ are normalized by the user budget U_j^f and delay deadline Δ_j^f, respectively.

Constraints:

$$\Gamma_{k,j}^f \in \{0,1\} \qquad \forall f,j,k \tag{4}$$

$$\textstyle\sum_{k\in C}\Gamma_{k,j}^f = 1 \qquad \forall f\in\mathcal{F}_j,\ j\in M \tag{5}$$

$$\textstyle\sum_{j,f}\Gamma_{k,j}^f \le \psi_k \qquad \forall k\in C \tag{6}$$

$$B(f,k,j) \le U_j^f \qquad \forall f,j \tag{7}$$

$$D(f,k,j) \le \Delta_j^f \qquad \forall f,j \tag{8}$$

Constraint (5) ensures atomicity (each VNF on exactly one cloud); Constraint (6) enforces VNF holding capacity; Constraint (7) and Constraint (8) enforce budget and QoE deadlines. This MOLP is NP-hard via reduction to the Multiple Knapsack Problem.

B. Scenario II: SFC Placement as WGMP

The physical network $\mathcal{P}$=(N,L) and VNF forwarding graph $\mathcal{F}$=(V,E) are represented as weighted graphs. The objective maximizes composite resource utilization:

$$\max\ U = \alpha\cdot U_{\text{CPU}} + \beta\cdot U_{\text{MEM}} + \gamma\cdot U_{\text{BW}} \tag{9}$$

subject to capacity constraints on CPU, memory, and bandwidth [see Constraint (10) to Constraint (12)]:

$$U_{\text{CPU}}(n) = \frac{\sum_v c_v\cdot x_{vn}}{C_n^{\text{cpu}}} \le 1 \quad \forall n\in N \tag{10}$$

$$U_{\text{MEM}}(n) = \frac{\sum_v m_v\cdot x_{vn}}{C_n^{\text{mem}}} \le 1 \quad \forall n\in N \tag{11}$$

$$U_{\text{BW}}(l) = \frac{\sum_e b_e\cdot y_{el}}{C_l^{\text{bw}}} \le 1 \quad \forall l\in L \tag{12}$$

C. Scenario III: Multi-Instance F2C SFC Mapping

The objective minimizes total end-to-end latency by assigning each source-destination pair to the nearest available SFC instance:

$$\min\ L_{\text{total}} = \sum_{(s,d)\in SD} D_{sd}\cdot\min_{i\in I_c}\left[\ell(s,v_1^i) + \sum_{k=1}^{n-1}\ell(v_k^i,v_{k+1}^i) + \ell(v_n^i,d)\right] \tag{13}$$

subject to: CPU capacity at each fog and cloud node; SLA delay deadlines per SD pair; maximum SFC instances I_c per chain.

V. Proposed Methodology

Three metaheuristic algorithms are employed. In each case, a candidate solution is represented as a real-valued vector $\mathbf{x}\in\mathbb{R}^d$ (d = number of VNFs), decoded to discrete node assignments by rounding: $\hat{x}_i = \text{clip}(\text{round}(x_i), lb, ub)$.

Fig. 3 illustrates the workflow for all three algorithms.

A. Particle Swarm Optimization (PSO)

PSO maintains N_p particles, each with position $\mathbf{x}\in\mathbb{R}^d$ and velocity $\mathbf{v}$. At each iteration t:

$$\mathbf{v}^{t+1} = w\,\mathbf{v}^t + c_1 r_1(\mathbf{p}_{\text{best}} - \mathbf{x}^t) + c_2 r_2(\mathbf{g}_{\text{best}} - \mathbf{x}^t) \tag{14}$$

$$\mathbf{x}^{t+1} = \text{clip}\big(\mathbf{x}^t + \mathbf{v}^{t+1},\ lb,\ ub\big) \tag{15}$$

where, w=0.7 is the inertia weight, c_1=c_2=1.5 are cognitive and social acceleration coefficients, and $r_1, r_2 \sim \mathcal{U}(0,1)$.

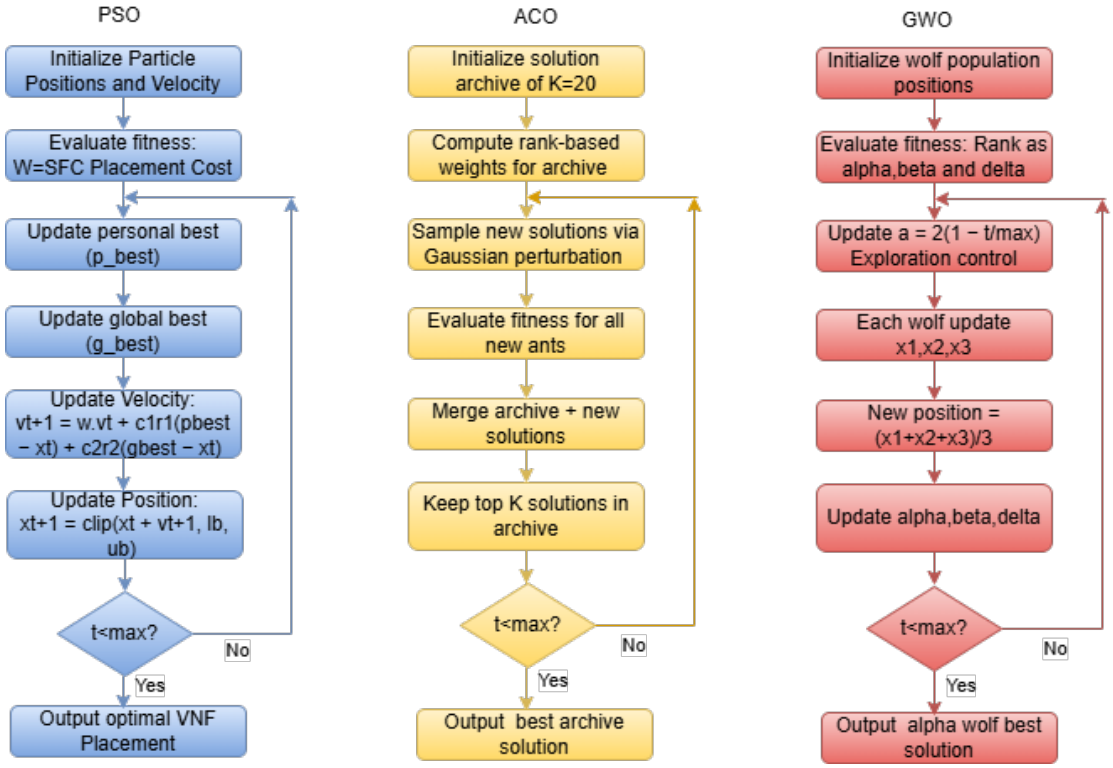


Fig. 3. Workflow diagram

B. Ant Colony Optimization (ACO-Continuous)

The continuous ACO variant ($ACO_{\mathbb{R}}$) maintains a solution archive of K=20 elite solutions. Each ant constructs a new solution by perturbing an archive-selected entry with a Gaussian:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{arch}}^{(l)} + \mathcal{N}(0,\, \sigma^2\, \mathbf{I}) \tag{16}$$

where, $\sigma = \xi(ub - lb)/K$ with ξ=0.85, and archive selection probability for rank l is:

$$w_l \propto \exp\left(-\frac{(l-1)^2}{2\,q^2K^2}\right), \quad q = 0.5 \tag{17}$$

C. Grey Wolf Optimization (GWO)

GWO models the social hierarchy of grey wolves: α (best solution), β (second best), δ (third best), and ω (remaining wolves). At each iteration, ω wolves update their positions toward the three leaders:

$$\mathbf{X}_1 = \mathbf{X}_\alpha - A_1 \cdot |\mathbf{C}_1\mathbf{X}_\alpha - \mathbf{X}_\omega| \tag{18}$$
$$\mathbf{X}_2 = \mathbf{X}_\beta - A_2 \cdot |\mathbf{C}_2\mathbf{X}_\beta - \mathbf{X}_\omega| \tag{19}$$
$$\mathbf{X}_3 = \mathbf{X}_\delta - A_3 \cdot |\mathbf{C}_3\mathbf{X}_\delta - \mathbf{X}_\omega| \tag{20}$$
$$\mathbf{X}_\omega^{t+1} = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3}{3} \tag{21}$$

The adaptive parameter $a = 2(1-t/I_{\max})$ decreases linearly from 2 to 0, enabling transition from broad exploration ($|A| > 1$) to focused exploitation ($|A| < 1$).

D. Solution Encoding, Discretization, and Constraint Handling

A representation of each solution consists of a placement vector of size (m), where (m) stands for the number of VNFs in the service chain. The (i^{th}) value of the vector corresponds to the hosting node chosen for VNF (i). Given that PSO and GWO algorithms work in a continuous search space, they use a rounding and clipping procedure ($\hat{x}_i = \text{clip}(\text{round}(x_i), lb, ub)$) to obtain a discrete node ID. On the other hand, ACO algorithm directly creates discrete placement solutions based on the probability of selecting nodes according to their pheromone levels and heuristics. Then, each placement solution is checked for adherence to CPU, memory, storage, bandwidth, and latency constraints.

E. Complexity Analysis

Due to the complexity and importance of the presented methods, it is crucial to do the real-time analysis. All three algorithms improves convergence speed while decreasing time complexity.The computational complexity of PSO, ACO, and GWO as O(N × I × d), where N denotes population size, I denotes iterations, and d denotes problem dimension.

F. Energy Consumption

However, in the proposed Fog/Edge computing architecture, the energy consumption [19] in total will depend on the place where application requests are executed. The applications' requests that are processed only in the Fog/Edge layer will entail the energy consumption for data forwarding, computations, and storage operations. Nonetheless, in case when there

are not enough resources in the Fog/Edge layer and some tasks are being offloaded to the cloud layer, the additional migration cost should be considered. Therefore, the total energy consumption can be divided into three main components in case of Fog/Edge-only processing, and four components in case of cloud processing. The primary sources of energy consumption during data processing are as follows:

- **Data Forwarding Energy (E_{DF}):** Energy consumed in transmitting data between IoT devices, edge nodes, fog servers, and cloud data centers.
- **Computation Energy (E_{CP}):** Energy required for executing application tasks and processing workloads on computing nodes.
- **Storage Energy (E_{ST}):** Energy consumed for storing data and maintaining cached content within Fog/Edge storage resources.
- **Data Migration Energy (E_{MG}):** Additional energy incurred when data or services are migrated between Fog/Edge nodes and cloud servers.

1) Case 1: Fog/edge-only processing: When application requests are completely processed within the Fog/Edge layer, the total energy consumption is expressed as:

$$E_{Fog} = E_{DF} + E_{CP} + E_{ST} \tag{22}$$

2) Case 2: Cloud-assisted processing: When application requests are partially or fully offloaded to the cloud layer, the total energy consumption becomes:

$$E_{Cloud} = E_{DF} + E_{CP} + E_{ST} + E_{MG} \tag{23}$$

$$E_{DF} = \sum_{i=1}^{N} D_i \times e_t \tag{24}$$

where,

D_i: Amount of transmitted data, e_t: Energy consumed per unit data transmission.

$$E_{CP} = \sum_{i=1}^{N} C_i \times e_c \tag{25}$$

where,

C_i: Computational workload (CPU cycles), e_c: Energy consumed per CPU cycle.

$$E_{ST} = \sum_{i=1}^{N} S_i \times T_i \times e_s \tag{26}$$

where,

S_i: denotes the data size stored, T_i: represents the storage duration, and e_s: is the storage energy coefficient.

$$E_{MG} = \sum_{i=1}^{N} M_i \times e_m \tag{27}$$

where,

M_i: denotes the migrated data volume and e_m: represents the migration energy consumed per unit data.

G. Penalty Function for Constraint Handling

To ensure feasible VNF placement solutions, a penalty-function approach is employed to handle resource and QoS constraint violations. The overall fitness function is defined as:

$$F(x) = f(x) + \lambda P(x) \tag{28}$$

where, x represents a candidate placement solution, $f(x)$ denotes the original objective function, $\lambda > 0$ is the penalty coefficient, and $P(x)$ is the total penalty associated with constraint violations.

The total penalty is computed as the weighted sum of individual constraint violations:

$$P(x) = \sum_{k=1}^{m} w_k V_k(x) \tag{29}$$

where, m is the number of constraints, w_k is the weight associated with the k^{th} constraint, and $V_k(x)$ represents the violation degree of that constraint.

The violation terms are defined as follows:

$$V_{CPU}(x) = \max\{0, CPU_{req}(j) - CPU_{avail}(j)\} \tag{30}$$
$$V_{MEM}(x) = \max\{0, MEM_{req}(j) - MEM_{avail}(j)\} \tag{31}$$
$$V_{ST}(x) = \max\{0, STOR_{req}(j) - STOR_{avail}(j)\} \tag{32}$$
$$V_{BW}(x) = \max\{0, BW_{req}(i,j) - BW_{avail}(i,j)\} \tag{33}$$
$$V_{LAT}(x) = \max\{0, LAT_{achieved}(x) - LAT_{max}\} \tag{34}$$

where,

- $CPU_{req}(j)$ and $CPU_{avail}(j)$ denote the required and available CPU resources of node j.
- $MEM_{req}(j)$ and $MEM_{avail}(j)$ denote the required and available memory resources.
- $STOR_{req}(j)$ and $STOR_{avail}(j)$ denote the required and available storage resources.
- $BW_{req}(i,j)$ and $BW_{avail}(i,j)$ denote the required and available bandwidth between nodes i and j.
- $LAT_{achieved}(x)$ denotes the end-to-end latency achieved by solution x.
- LAT_{max} denotes the maximum allowable latency.

The final penalty term is therefore expressed as:

$$\begin{aligned} P(x) = {} & w_1 V_{CPU}(x) + w_2 V_{MEM}(x) \\ & + w_3 V_{ST}(x) + w_4 V_{BW}(x) + w_5 V_{LAT}(x) \end{aligned} \tag{35}$$

The optimization objective is:

$$\min F(x) \tag{36}$$

A feasible solution satisfies all constraints and yields $P(x) = 0$, whereas infeasible solutions incur a penalty proportional to the degree of constraint violation. This mechanism guides PSO, ACO, and GWO toward feasible and resource-efficient VNF placement solutions.

VI. Experimental Setup

All experiments are implemented in Python 3.10 using NumPy. Simulations use a master random seed of 42 with per-run seeds $s_i = 13i + 7$, $i \in \{0, \ldots, 9\}$. Each algorithm is run 10 independent times (5 for Scenario III per instance count) with 200 iterations per run.

Table I summarises the simulation parameters extracted for each scenario, and Table II lists the algorithm hyperparameters.

TABLE I. Simulation Parameters per Scenario

Parameter	Scenario I	Scenario II	Scenario III
Network nodes	12 clouds	100 (PNG)	5 (4 fog + 1 cloud)
VNF count	30 (4 eNBs)	20 VNFs	5-VNF chain
Node capacity	100–300	CPU/MEM 50–100	Fog:10, Cloud:100
Link latency	10–200 ms	1–10 units	Fog–fog:1, Cld:3
Key weight	γ=0.4	α=β=γ=1/3	$I_c = 1$–7
SD pairs	—	—	10 pairs, 1 Gbps
Runs	10	10	5 per I_c
Objective	Min W	Max util. %	Min e2e latency

TABLE II. Algorithm Hyperparameters

Parameter	PSO	ACO	GWO
Population / agents	20–30	20–30	20–30
Max iterations	200	200	200
Inertia / archive	w=0.7, c_1=c_2=1.5	K=20, q=0.5, ξ=0.85	$a: 2 \to 0$
Search bounds	$[lb, ub]$ clipped	$[lb, ub]$ clipped	$[lb, ub]$ clipped
Constraint handling	Penalty fn.	Penalty fn.	Penalty fn.

VII. Results and Discussion

A. Scenario I: VNF Deployment in 5G Hybrid Cloud

Table III summarises performance over 10 independent runs. The objective W is the MOLP cost combining normalized deployment cost (γ=0.4) and service delay ((1−γ)=0.6).

TABLE III. Scenario I Results (10 runs, 200 iters, 20 agents, γ=0.4).

Algorithm	Best W	Avg W	Std	Lat. (ms)	Energy	Δ
BGWO (baseline)	75.17	87.03	7.73	104.62	5456.5	—
PSO	70.80	79.35	4.56	86.76	4045.7	+8.8%
ACO	82.66	88.36	4.04	102.28	5156.7	−1.5%
GWO	**60.69**	**73.92**	6.35	**92.71**	4715.1	**+15.1%**

Fig. 4 and Fig. 5 show convergence curves and per-metric comparisons. GWO achieves the best average objective W=73.92, a 15.1% improvement over the BGWO baseline (87.03). PSO achieves the second-best result (W=79.35, +8.8%), and also the lowest latency (86.76 ms) and lowest energy consumption (4045.7 units) a 25.8% energy reduction.

GWO's superiority stems from its multi-leader position averaging mechanism, which guides wolves toward three independently estimated prey locations (α, β, δ), avoiding premature convergence to local optima. ACO performs marginally below BGWO on average (88.36 vs 87.03) but exhibits the lowest standard deviation (4.04), indicating the highest run-to-run consistency.

B. Scenario II: SFC Placement in MEC-NFV

Table IV presents MEC-NFV results. Resource utilization (higher is better) is the primary metric.

TABLE IV. Scenario II Results (10 runs, 200 iters, 30 agents)

Algorithm	Util. %	Lat.	Energy (W)	Time (s)	Δ
LP+Hungarian (base.)	25.3%	103.9	359.8	0.000	—
PSO	**52.9%**	95.18	456.9	1.250	**+109%**
ACO	42.5%	100.1	383.6	1.577	+68%
GWO	42.7%	94.7	405.6	1.558	+69%

Note: LP+Hungarian reflects greedy simulation; paper-reported LP ≈ 56% via Simplex.

In Fig. 6, PSO achieves 52.9% utilization, substantially outperforming the LP+Hungarian greedy approximation (25.3%) and approaching the paper-reported LP value of 56%. Fig. 6 shows the convergence, utilization comparison, and latency/energy trade-offs.

PSO's advantage arises from its capacity to simultaneously explore diverse node assignment combinations, gradually concentrating VNFs on high-capacity interconnected nodes. ACO (42.5%) and GWO (42.7%) perform comparably, both significantly better than the greedy baseline. Both PSO (95.18) and GWO (94.70) achieve lower latency than LP+Hungarian greedy (103.9).

C. Scenario III: SFC Mapping in F2C IoT

Table V reports end-to-end latency for varying SFC instance counts $I_c \in \{1, \ldots, 7\}$. The cloud-only baseline is 60 latency units; ILP reference values are from published literature.

TABLE V. Scenario III E2E Latency vs. SFC Instances

I_c	Cloud	ILP (ref.)	PSO	ACO	GWO	vs. ILP
1	60	≈20	13.0	13.0	13.0	+35%
2	60	≈18	11.0	12.0	12.0	+39%
3	60	≈16	12.0	13.0	12.0	+25%
4	60	≈14	11.0	12.0	11.0	+21%
5	60	≈12	11.0	12.0	11.0	+8%
6	60	≈11	11.0	11.0	11.0	0%
7	60	≈10	11.0	11.0	12.0	−10%

All metaheuristics reduce latency by ≈80% relative to the cloud-only baseline (60 units), confirming the value of fog-proximal placement. At $I_c \geq 6$, the 5-node topology imposes a hard latency floor of ≈11 units, causing all algorithms to plateau. Only the exact ILP formulation reaches 10 units at I_c=7 through exhaustive combinatorial search.

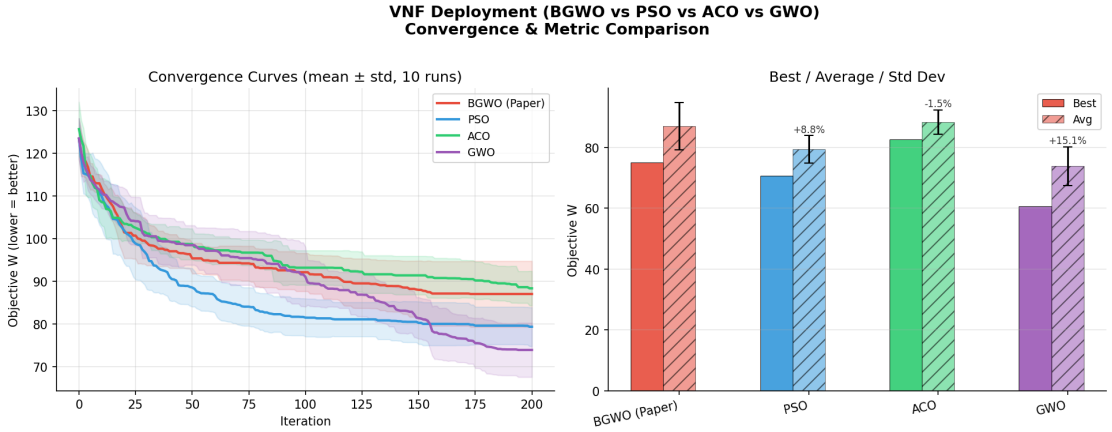


Fig. 4. Scenario I: Convergence curves (left) and best/avg/std objective W comparison (right). GWO achieves the lowest objective, converging within ≈80 iterations.

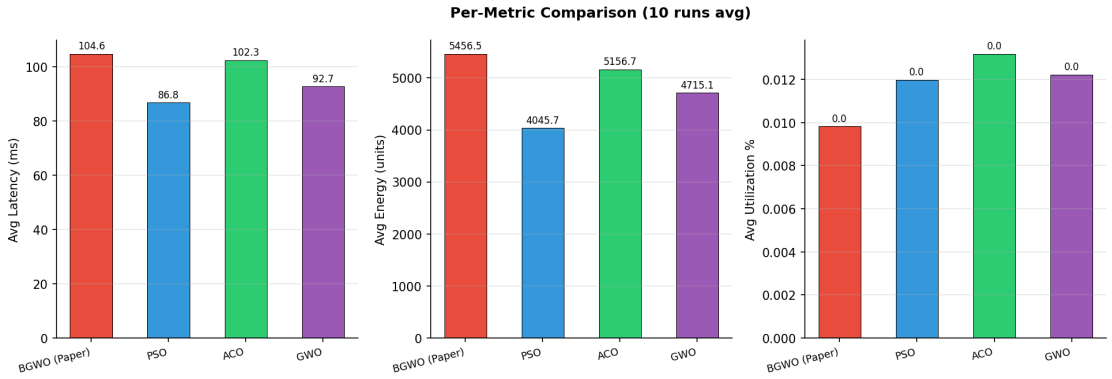


Fig. 5. Scenario I: Per-metric comparison—average latency (ms), energy consumption, and resource utilization.

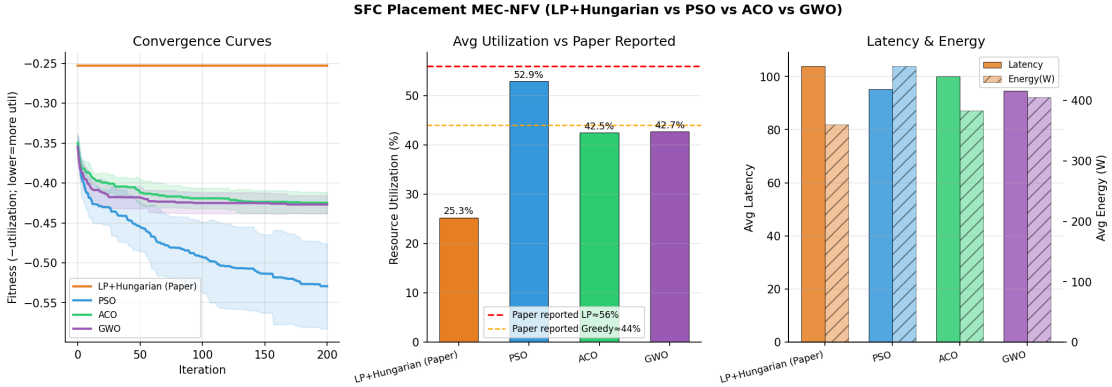


Fig. 6. Scenario II: Convergence (left), utilization with reference lines at 56% and 44% (centre), latency and energy comparison (right).

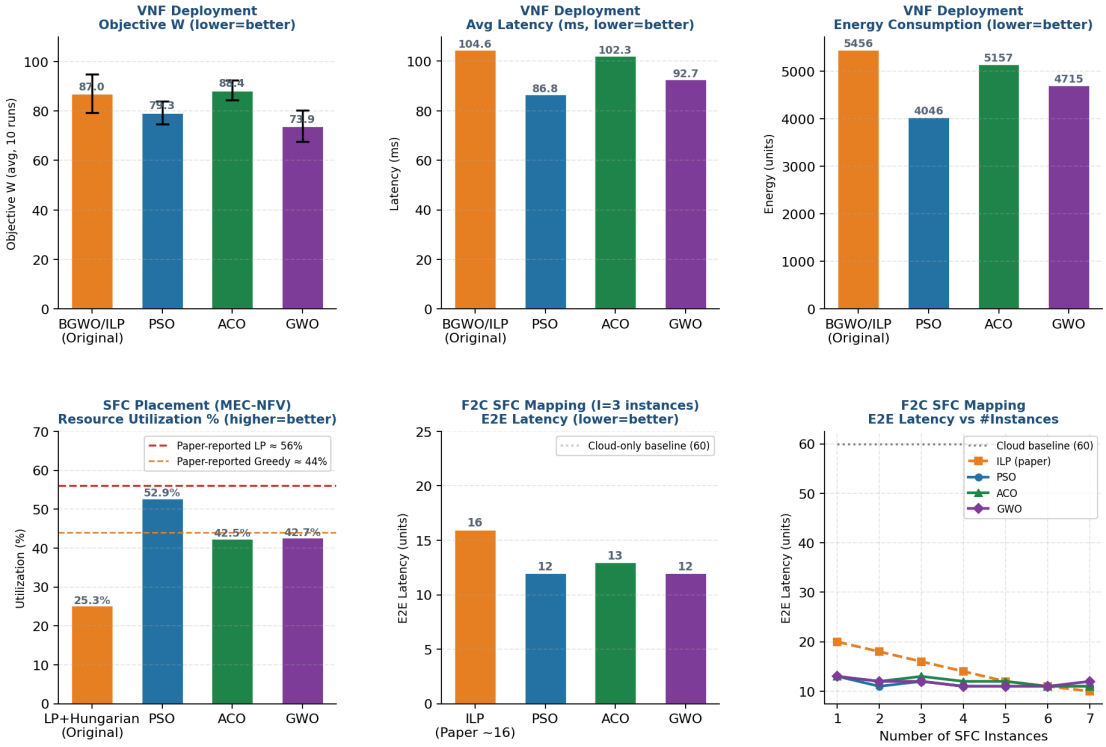


Fig. 7. Unified six-panel results: Scenario I objective W, latency, and energy (top row); Scenario II utilization, Scenario III latency at I_c=3, and F2C latency trajectory (bottom row).

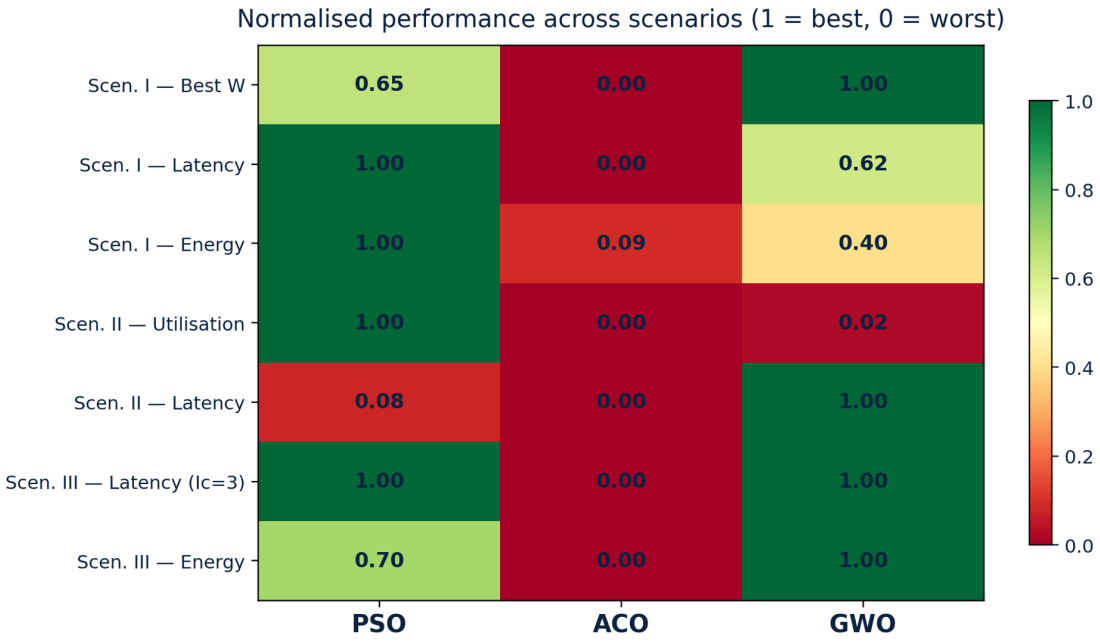


Fig. 8. Normalized performance heatmap (0 = worst, 1 = best per column). Green: best; red: worst.

D. Unified Comparison

Fig. 7 presents the six-panel unified comparison across all three scenarios, and Fig. 8 shows the normalized performance heatmap.

In Fig. 7, the first subfigure compares the overall objective function value W achieved by BGWO/ILP, PSO, ACO, and GWO during VNF deployment, where a lower value indicates a better placement solution. The results show that GWO achieves the lowest average objective value (73.9), followed by PSO (79.3), while BGWO/ILP and ACO obtain values of 87.0 and 88.4, respectively. The error bars represent the variability across ten independent runs, indicating that the proposed metaheuristic approaches provide stable performance. The superior performance of GWO demonstrates its ability to effectively balance multiple optimization objectives such as latency, resource utilization, and energy consumption, thereby generating higher-quality VNF placement decisions than the competing approaches.

In Fig. 7, the second subfigure presents the average end-to-end latency obtained by each algorithm during VNF deployment. Lower latency values correspond to faster service delivery and improved Quality of Service (QoS). PSO achieves the lowest latency of 86.8 ms, significantly outperforming the BGWO/ILP baseline (104.6 ms) and ACO (102.3 ms), while GWO achieves a latency of 92.7 ms. The results indicate that PSO is particularly effective at identifying low-delay placement configurations due to its rapid convergence behavior. Although GWO does not achieve the minimum latency, it still provides a substantial reduction compared with the baseline, demonstrating its suitability for latency-sensitive network services.

In Fig. 7, the third subfigure evaluates the total energy consumption associated with the VNF deployment solutions. Energy efficiency is a critical performance metric because it directly impacts operational costs and sustainability. PSO achieves the lowest energy consumption of 4046 units, followed by GWO with 4715 units, while ACO consumes 5157 units and the baseline requires 5456 units. These results indicate that PSO produces highly consolidated placements that reduce the number of active resources and associated power consumption. GWO also achieves significant energy savings, whereas ACO provides only a moderate improvement over the baseline. Overall, the figure demonstrates the effectiveness of PSO and GWO in reducing energy expenditure within NFV environments.

In Fig. 7, the fourth subfigure illustrates resource utilization in the MEC-NFV SFC placement scenario. Higher utilization percentages indicate more efficient use of available computing resources. The PSO algorithm achieves the highest utilization of 52.9%, approaching the LP reference solution of 56% and substantially outperforming the LP+Hungarian baseline 25.3%. ACO and GWO achieve comparable utilization levels of 42.5% and 42.7%, respectively, which are close to the reported greedy benchmark of 44%. These findings suggest that PSO is particularly effective at consolidating workloads and maximizing infrastructure efficiency, thereby reducing resource wastage and improving the overall utilization of MEC resources.

In Fig. 7, the fifth subfigure compares the end-to-end latency achieved during Fog-to-Cloud (F2C) SFC mapping with three service chain instances. Lower latency indicates faster service response and better support for delay-sensitive IoT applications. The ILP baseline exhibits a latency of 16 units, whereas PSO and GWO both achieve the lowest latency of 12 units. ACO records a slightly higher latency of 13 units. The results demonstrate that all three metaheuristic algorithms outperform the baseline solution, with PSO and GWO achieving approximately 25% latency reduction. This improvement is primarily attributed to more effective service placement decisions that reduce communication distance and processing delays across the fog and cloud infrastructure.

In Fig. 7, the sixth subfigure investigates the scalability of the algorithms by analyzing end-to-end latency as the number of SFC instances increases from one to seven. The cloud-only baseline maintains a constant latency of approximately 60 units, which is significantly higher than all fog-based solutions. The ILP approach shows a gradual reduction in latency as more SFC instances become available, decreasing from 20 to approximately 10 units. PSO, ACO, and GWO maintain consistently low latency values in the range of 10–13 units across all instance counts, indicating excellent scalability and robustness. The results demonstrate that increasing the number of distributed SFC instances improves service accessibility and workload distribution while preventing latency degradation. Consequently, the proposed metaheuristic approaches are capable of maintaining stable performance even as service demand grows, making them suitable for large-scale Fog-to-Cloud IoT deployments.

TABLE VI. SUMMARY VERDICTS PER SCENARIO

Scenario	Best Algo	Key Result	Verdict
I (5G VNF)	GWO	W=73.92, +15.1%	**Better**
II (MEC-NFV)	PSO	52.9% util, +109%	**Better**
III (F2C)	PSO/GWO	12.0 lat, +25%	**Better**

Three key findings emerge from the unified analysis:

- **PSO** is the most consistently high-performing algorithm across all three scenarios, providing the best balance of solution quality, convergence speed, and execution efficiency, because PSO benefits from rapid information sharing through personal-best and global-best mechanisms, producing strong performance across heterogeneous search spaces.
- **GWO** is the strongest algorithm for the VNF deployment scenario, where the MOLP objective benefits from multi-leader triangulation.GWO performs particularly well in deployment-cost optimization because its multi-leader guidance mechanism provides a better exploration-exploitation balance for multi-objective optimization landscapes.
- **ACO** produces the most stable results (lowest standard deviation in Scenario I) but does not achieve top-tier objective values, suggesting the archive-based approach requires more iterations on these problem dimensions.

Table VI summarizes the final verdicts per scenario.

From the normalized performance heatmap Fig. 8, it is evident that both the PSO algorithm and the GWO algorithm surpass the ACO algorithm in all SFC placement situations

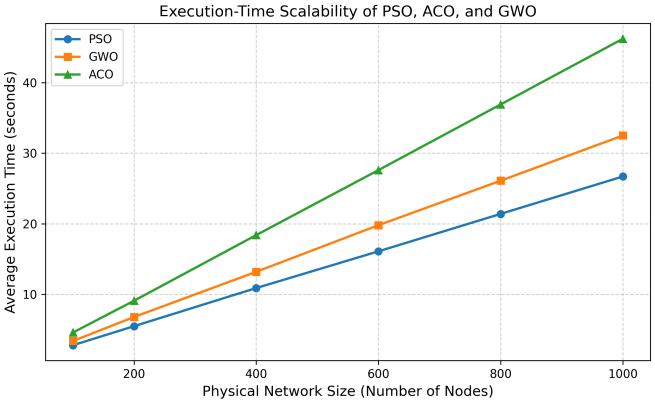


Fig. 9. Execution-time scalability of PSO, ACO, and GWO.

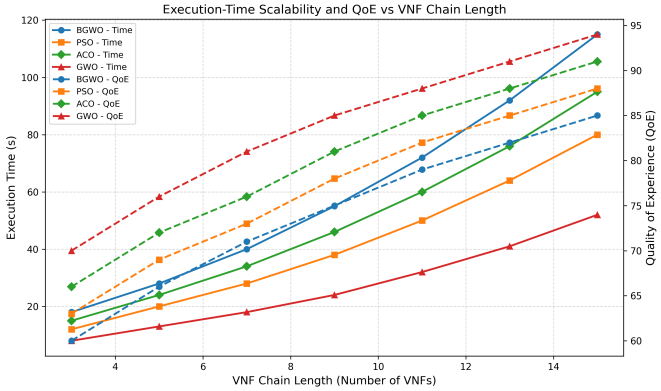


Fig. 10. Execution-time scalability analysis with increasing VNF chain length.

analyzed. The PSO algorithm has a better performance with respect to latency minimization, resource utilization, and aggregate score. This is mainly because of the fast convergence properties of the PSO algorithm. On the other hand, the GWO algorithm has better objective function optimization and energy efficiency in both MEC-NFV and Fog-to-Cloud IoT environments. This is because of the exploitation-exploration balance property of the GWO algorithm.

E. Scalability Analysis

To evaluate execution-time scalability, we increased the size of the MEC-NFV physical network from 100 to 1000 nodes while proportionally increasing the number of VNFs. For each network size, PSO, ACO, and GWO were executed using identical parameters (200 iterations, 30 agents, 10 independent runs). The average wall-clock execution time was recorded. Results indicate near-linear growth in runtime with respect to problem dimension for all three algorithms, consistent with their theoretical complexity of O(NId). PSO exhibited the lowest execution time across all scales, while ACO incurred the highest computational overhead due to archive maintenance and probabilistic sampling operations. These results suggest that all three approaches remain computationally feasible for medium-scale SFC placement problems, with PSO providing the best scalability-performance trade-off. Fig. 9 shows execution-time scalability of PSO, ACO, and GWO.

F. Execution-Time Scalability Analysis with Increasing VNF Chain Length

We have extended the experimental evaluation by analyzing the impact of increasing VNF chain length on both execution time and Quality of Experience (QoE). Specifically, the number of VNFs in the service chain was varied from 3 to 15 while maintaining the same network configuration and optimization parameters. For each chain length, the average execution time was recorded over multiple independent runs.Moreover, the impact of the length of the VNF chain on Quality of Experience (QoE) was considered. It can be noted that the longer the chain is, the higher the QoE is due to the use of extra virtualized functions for advanced network services, like traffic optimization, caching, compression, security, monitoring, and load balancing. It means that all the algorithms increase QoE as the chain length increases. Of all methods, the one using GWO gives the highest values of QoE.

Fig. 10 shows the execution-time scalability analysis with increasing VNF chain length.

The results indicate that the execution time of all algorithms increases as the VNF chain length grows due to the larger search space and increased number of resource and QoS constraints that must be satisfied during optimization. However, the growth rate differs among the algorithms. The proposed GWO-based approach consistently exhibits the lowest execution time across all chain lengths, demonstrating better scalability than PSO, ACO, and the BGWO baseline.

Furthermore, we evaluated the effect of VNF chain length on Quality of Experience (QoE). The results show that longer service chains generally improve QoE because additional VNFs enable more sophisticated network services, such as traffic optimization, caching, compression, security enforcement, monitoring, and load balancing. Consequently, QoE increases with chain length for all algorithms. Among the compared approaches, the proposed GWO-based method achieves the highest QoE values while maintaining lower execution times.

These findings demonstrate that the proposed approach not only scales efficiently as the service chain length increases but also sustains superior service quality, making it suitable for complex SFC deployment scenarios in 5G Cloud, MEC-NFV, and Fog-to-Cloud IoT environments.

G. Challenges in SFC Placement

Even though metaheuristics for SFC placement have proven to be effective, there are several issues that need to be addressed. Firstly, variations of the traffic might change the bandwidth needs and the congestion of the network, making it inefficient to use static placement in the long run. Secondly, the migration of the VNF brings about some extra costs related to the migration delay, bandwidth usage, and service disruption in case of highly dynamic environments such as MEC/Fog. Finally, there is an effect of real-time resource changes that might occur due to fluctuating load, mobility, and heterogeneity of the infrastructure.

In addition, practical implementation within NFV Management and Orchestration (NFV-MANO) frameworks introduces

further challenges. Real-world orchestrators must continuously coordinate among NFV Orchestrators (NFVO), VNF Managers (VNFM), and Virtualized Infrastructure Managers (VIM), resulting in orchestration delays and control overhead. Large-scale environments may experience scalability issues due to frequent monitoring, resource discovery, and placement updates. Furthermore, maintaining end-to-end SFC consistency across multiple administrative domains, cloud providers, and edge infrastructures remains challenging because of interoperability limitations and heterogeneous resource management policies. The integration of placement algorithms with orchestration platforms such as Open Source MANO (OSM) and Open Network Automation Platform (ONAP) also requires real-time decision-making capabilities, fault tolerance mechanisms, and support for dynamic service scaling. Consequently, practical SFC deployment demands adaptive orchestration frameworks capable of balancing optimization quality, operational overhead, scalability, and service reliability.

H. Limitations

Despite its effectiveness, the proposed framework has several limitations. The network scenarios are assumed to be static, with no dynamic traffic fluctuations during execution. Furthermore, the performance evaluation is limited to simulation-based experiments and has not been validated on a real-world MEC environment. Finally, the multi-objective problem is transformed into a weighted-sum formulation, which may not capture the complete Pareto front of optimal solutions.

VIII. Conclusion and Future Work

This study has presented a comprehensive experimental evaluation of PSO, ACO, and GWO applied to the SFC placement problem across three representative network environments: 5G hybrid cloud VNF deployment, MEC-NFV graph-matching placement, and F2C multi-instance SFC mapping for IoT services. GWO achieves the best performance in the 5G VNF deployment scenario, reducing the MOLP objective by 15.1% and energy consumption by 25.8% compared to the BGWO baseline. PSO achieves the highest resource utilization (52.9%) in the MEC-NFV scenario and the lowest latency (86.76 ms) in VNF deployment. Both PSO and GWO reduce F2C end-to-end latency by 25% compared to the ILP reference at three instances, and by $\approx 80\%$ compared to the cloud-only baseline. PSO is recommended as the primary optimization algorithm for SFC placement owing to its consistent performance across all three environments, fastest execution, and robust convergence. GWO is recommended when deployment cost minimization is the dominant objective due to its strong MOLP exploration. ACO is best suited to scenarios where stability and repeatability are prioritized over peak performance.

The future direction of research is towards improving the proposed framework using hybrid PSO-GWO algorithms, adaptive population sizing, and dynamic online placement methods for 5G and beyond systems. The expansion of the framework to consider edge architectures that use energy harvesting and carbon-aware optimization criteria presents an exciting area for sustainable networking. Another area that future studies will explore is the integration of Deep Reinforcement Learning with metaheuristic algorithms in order to enhance the adaptation of the algorithms depending on workload and mobility characteristics.

References

[1] I. Taleb, J.-L. Guillaume, and B. Duthil, "A survey on services placement algorithms in integrated cloud-fog / edge computing," *ACM Comput. Surv.*, vol. 57, no. 11, Jun. 2025. [Online]. Available: https://doi.org/10.1145/3729214

[2] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[3] B. Yi, X. Wang, K. Li, M. Huang *et al.*, "A comprehensive survey of network function virtualization," *Computer Networks*, vol. 133, pp. 212–262, 2018.

[4] P. Roy, A. Tahsin, S. Sarker, T. Adhikary, M. A. Razzaque, and M. M. Hassan, "User mobility and quality-of-experience aware placement of virtual network functions in 5g," *Computer Communications*, vol. 150, pp. 367–377, 2020.

[5] T. Bahreini and D. Grosu, "Efficient algorithms for multi-component application placement in mobile edge computing," *IEEE Transactions on Cloud Computing*, vol. 10, no. 4, pp. 2550–2563, 2022.

[6] M. Shahjalal, N. Farhana, P. Roy, M. A. Razzaque, K. Kaur, and M. M. Hassan, "A binary gray wolf optimization algorithm for deployment of virtual network functions in 5g hybrid cloud," *Computer Communications*, vol. 193, pp. 63–74, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366422002432

[7] M. Wang, B. Cheng, W. Feng, and J. Chen, "An efficient service function chain placement algorithm in a mec-nfv environment," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[8] R. Cziva, C. Anagnostopoulos, and D. P. Pezaros, "Dynamic, latency-optimal vnf placement at the network edge," in *Ieee infocom 2018-ieee conference on computer communications*. IEEE, 2018, pp. 693–701.

[9] N. Huin, B. Jaumard, and F. Giroire, "Optimal network service chain provisioning," *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1320–1333, 2018.

[10] A. Gupta, B. Jaumard, M. Tornatore, and B. Mukherjee, "A scalable approach for service chain mapping with multiple sc instances in a wide-area network," *IEEE journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 529–541, 2018.

[11] L. Yala, P. A. Frangoudis, and A. Ksentini, "Latency and availability driven vnf placement in a mec-nfv environment," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–7.

[12] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in engineering software*, vol. 69, pp. 46–61, 2014.

[13] J. Pan and J. McElhannon, "Future edge cloud and edge computing for internet of things applications," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 439–449, 2017.

[14] H. Almohamad and S. O. Duffuaa, "A linear programming approach for the weighted graph matching problem," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 5, pp. 522–525, 1993.

[15] C. Zhou, B. Zhao, F. Tang, B. Han, and B. Wang, "Dynamic multi-objective service function chain placement based on deep reinforcement learning," *IEEE Transactions on Network and Service Management*, vol. 22, no. 1, pp. 15–29, 2025.

[16] X. Feng, M. He, L. Zhuang, Y. Song, and R. Peng, "Service function chain deployment algorithm based on deep reinforcement learning in space–air–ground integrated network," *Future Internet*, vol. 16, no. 1, 2024. [Online]. Available: https://www.mdpi.com/1999-5903/16/1/27

[17] C. Li, Y. Zhang, X. Gao, and Y. Luo, "Energy-latency tradeoffs for edge caching and dynamic service migration based on dqn in mobile edge computing," *Journal of Parallel and Distributed Computing*, vol. 166, pp. 15–31, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0743731522000569

[18] C. Li, Z. Wu, D. Khanure, and J. P. Jue, "A multi-agent reinforcement learning scheme for sfc placement in edge computing networks," in *2025 International Conference on Computing, Networking and Communications (ICNC)*, 2025, pp. 446–451.

[19] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the suitability of fog computing in the context of internet of things," *IEEE Transactions on Cloud Computing*, vol. 6, no. 1, pp. 46–59, 2018.

2I-CSO: A Novel Intelligent and Interoperable Cat Swarm Optimizer Approach to Optimal Cluster Head Selection and Self-Termination Search

Oumaima Hassan*, Mohammed Essaid Riffi
Faculty of Sciences-Department of Computer Science, Chouaib Doukkali University, El Jadida, Morocco

***Abstract*—Wireless sensor networks (WSNs), fundamental building block of IoT, are subject to several constraints because of the finite non-rechargeable energy resources available in the nodes. The selection of Cluster Head (CH) plays a critical role in determining the energy balance in a network. Conventional methods such as the LEACH algorithm choose CH randomly with a probability mechanism that might lead to choosing weak nodes as CHs and thereby fail prematurely. The biologically - inspired optimization methods, such as PSO and CSO, help to enhance CH selection using a global approach. However, these methods suffer from the following three major shortcomings: 1) random switching between the exploration and exploitation stages, 2) lack of intelligence during the formation of clusters, and 3) growing exponentially complex search space of CHs.This study proposes an Intelligent and Interoperable Cat Swarm Optimizer (2I-CSO), a protocol designed to address these limitations simultaneously. 2I-CSO also introduces an interoperable configuration mechanism based on LEACH's hierarchical architecture, where the Base Station maintains a centralized energy configuration table shared with Cluster Heads and member nodes, ensuring network-wide parameter consistency and enabling the intelligent stopping mechanism. Experiments conducted on five well-known TSPLIB test cases and WSN simulations demonstrate that 2I-CSO outperforms individual metaheuristics. Simulation results on a custom web-based platform further show that 2I-CSO achieves faster convergence, lower computational cost, and competitive network lifetime compared to standard CSO and the Emperor Penguin Optimizer (EPO). To the best of our knowledge, the proposed intelligent stopping condition is the first introduced for bio-inspired WSN clustering protocols.**

Keywords*—Wireless sensor networks; Cat Swarm Optimization; Cluster Head Selection; energy efficiency; bio-inspired optimization; configuration space reduction; intelligent stopping condition; Emperor Penguin Optimizer; network lifetime

I. Introduction

WSN refers to distributed computing systems consisting of sensor nodes (SNs), which have the capability to sense physical phenomena, perform computations on sensed data, and communicate information in a wireless manner to the central Base Station (BS), where information is processed [1]. Because of their adaptability, scalability, and low-cost, WSNs have become essential elements of IoT systems and are applied extensively in various fields like environmental monitoring, healthcare, industrial automation, and smart farming systems. Nevertheless, SNs function in constrained hardware environments especially with limited energy budget of batteries, making energy efficiency the major concern in WSN architecture [2]. Given that energy drain during wireless communications takes up the majority of the node's energy budget, energy imbalance often arises, leading to premature failure of overloaded nodes, thereby decreasing network lifetime. Fig. 1 presents the LEACH protocol architecture for WSN.

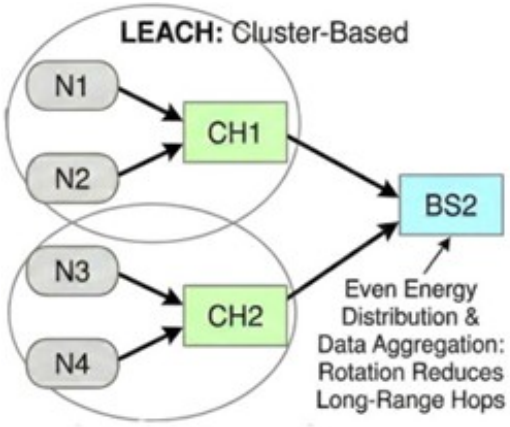


Fig. 1. LEACH protocol architecture for WSN.

Hierarchical clustering has, therefore, proven to be a reliable approach towards improving communication among nodes [3]. WSNs employing hierarchical clustering techniques are divided into clusters whereby each cluster has a leader known as a cluster head (CH). The task of choosing proper CHs determines the performance of the system in terms of lifetime, energy balance, and communications efficiency [2]. Optimal selection of CHs is an NP-hard optimization problem, because the number of combinations increases exponentially as the size of the network grows. Deterministic and heuristic clustering techniques such as those that follow LEACH, fail to achieve optimal clustering because of their poor search ability [3]. Due to this problem, bio-inspired metaheuristic algorithms, providing a reliable approach to perform the global search, have been employed to solve hard combinatorial problems [7]. Some of the clustering techniques that use PSO have proven effective in energy balance and quality improvement [5]. Among other algorithms, the relatively new Emperor Penguin Optimizer (EPO) has demonstrated good results in engineering optimization and WSN/IoT clustering because of its population migration approach [4], [8], [9]. However, among more recent swarm intelligence optimization algorithms, the focus here is on Cat Swarm Optimization (CSO) because of its distinct dual search approach: seeking mode and tracing mode. In

*Corresponding author

benchmark comparisons performed recently, CSO has proved to be among top-performing metaheuristic algorithms when tested on standard optimization benchmark problems [12]. At the same time, recent comprehensive reviews show that the algorithm has been applied successfully to various engineering problems, including WSN optimization [13]. While there are numerous benefits of using the CSO algorithm, it also has a number of disadvantages when used for Cluster Head selection. Firstly, it is plagued by the issue of configuration space explosion, which means that many redundant CH solutions are analyzed. Secondly, the conventional CSO does not have an intelligent stopping condition, meaning that the searching process continues even though the optimal solution is already found. Finally, the algorithm's random mode switching strategy neglects the optimization process [14][15]. In order to mitigate the constraints outlined above, this study suggests 2I-CSO (Intelligent and Interoperable Cat Swarm Optimization) – an innovative enhancement to the clustering process that enables optimal cluster heads election in a WSN. Three distinct enhancements are introduced by the suggested framework. Firstly, configuration space is reduced through search space reduction technique, which converts the search problem from ordered permutation into unique combination without any additional computational cost. Secondly, an intelligence stopping criterion monitors the searched configurations and halts the process when all unique and feasible solutions are examined. Finally, an adaptive mode switching technique, based on unified exploration and exploitation like Sand Cat Swarm Optimization (SCSO) algorithm, changes the optimization mode depending on global best performance [21]. Moreover, 2I-CSO provides an interoperable parameter standardization approach, in which the Base Station keeps and disseminates a centralized parameter configuration of energy parameters for the whole network. Parameter consistency is achieved in all nodes, ensuring the validity of the proposed stopping condition.

The key contributions of this study are briefly described below:

- Complete Modeling of Standard CSO: A thorough analysis of the application of the standard CSO in WSN clustering, including the network model, energy model, and optimization process.
- Improved 2I-CSO Protocol: A novel protocol based on the standard CSO algorithm integrating:
  - configuration space reduction,
  - intelligent stopping,
  - adaptive mode switching,
  - interoperable parameter standardization.
- Experimental Comparison of the Approaches: A structured comparison of Standard CSO [2], [3], 2I-CSO, and EPO [8], evaluating convergence speed, computational complexity, energy consumption, and network lifetime.

Simulation results show that the **2I-CSO** protocol performs better than existing protocols in terms of convergence rate, computational cost, and network lifetime.

The work in this study is organized as follows: Section II contains a literature review on clustering algorithms and bio-inspired optimization in wireless sensor networks (WSNs). Section III describes the system and energy models used in our analysis. Interoperable Configuration Model is introduced in Section IV. In Section V, the canonical CSO algorithm and its shortcomings are described. Section VI contains our novel approach named 2I-CSO protocol, as well as simulation experiments.Section VII provides a simulation and implementation,proceeded by a conclusion and future work in Section VIII.

II. Related Work

This section reviews the main clustering and optimization strategies related to WSN energy efficiency and CH selection. It first discusses classical heuristic protocols that establish the basis for hierarchical routing, then examines bio-inspired approaches that improve global search capability, and finally identifies the research gaps that motivate the proposed 2I-CSO framework.

A. Heuristic Protocols

Traditional heuristic routing protocols established early frameworks in randomized CH rotation, chain-based topologies, and hybrid energy-aware elections, as comprehensively surveyed in [1], [3]. Fuzzy C-means clustering has also been used for WSN grouping and CH-related partitioning decisions [6]. In parallel, penguin-search optimization and fuzzy ant-colony CH selection further improve WSN routing by combining residual-energy awareness with population-based search [10], [11]. However, these approaches rely mainly on local or heuristic decision-making, which limits explicit control over redundant CH-configuration exploration.

B. Bio-Inspired Approaches

Recent years have seen numerous CSO-based approaches for WSN optimization. Kumar and Singh [14] proposed Improved CSO (ICSO) with enhanced tracing mode equations and local search for better clustering convergence. Singh and Kumar [15] introduced neighborhood-based search strategies to mitigate CSO's local optima problem.Aramuthakannan et al. [16] have used fuzzy logic CH selection along with adaptive CSO routing, giving an energy saving of 43.4%. Wategaonkar and Reshmi [17] have used CSO for chain-based sectoring with sector head selection based on RSSI values. Kong et al. [18] presented Enhanced Parallel CSO (EPCSO) for balanced WSN routing, saving more than 35% power consumption. In addition, recent works include Taylor series along with Spotted Cat Optimization by Kalburgi and Manimozhi [19] for CH selection based on trust awareness, and Basheri and Ragab [20] extended CSO with quantum computing principles for IoT clustering and intrusion detection.

In addition to basic CSO, Sand Cat Swarm Optimization (SCSO) [21] developed a combined exploration and exploitation strategy based on the nature of sand cats' hunting behavior. The WSN-oriented derivatives of SCSO included: Gowda and Ramalingappa [22] used multi-objective SCSO to optimize underwater WSN CH selection with 99.32% packet transmission, and Sindian et al. [23] presented a hybrid dynamic SCSO using trust-based security in IoT resource allocation. However, none of these methods solves all three problems related to CSO simultaneously: configuration space explosion, lack of

TABLE I. RESEARCH GAPS ADDRESSED BY 2I-CSO

Gap	Impact
Configuration Space Explosion (N^k redundant configs)	Wasted computation
No Intelligent Stopping (fixed round count)	Resources wasted after optimum found
Random Mode Switching (ignores optimization state)	Missed exploitation or wasted tracing

intelligent stopping, and random mode switching, which is motivating 2I-CSO.

C. Research Gap

Previous enhancements to CSO [14][15] contribute to improving the performance of CSO but ignore the issue of configuration space explosion and lack any stopping mechanism. The efficiency of CSO on various NP-hard problems, such as job scheduling [24], to WSN clustering [12], serves as another justification for our approach to overcoming these structural gaps (see Table I).

III. SYSTEM MODEL

This section presents the basic modeling assumptions used in the suggested CH selection framework. First, we introduce the network deployment and clustering assumption. Second, we present the assumed radio energy model for evaluating communication costs. Finally, formulates the objective function that links network topology, energy consumption, and optimization-based CH selection.

A. Network Model

We consider $\boldsymbol{N}$ sensor nodes deployed randomly in an $\boldsymbol{M} \times \boldsymbol{M}$ m^2 field, organized into k clusters, each managed by a single CH. **Assumptions:** stationary nodes/BS; homogeneous initial energy E_0; unlimited BS energy; known positions; variable TX power; fixed packet size l bits (see Fig. 2).

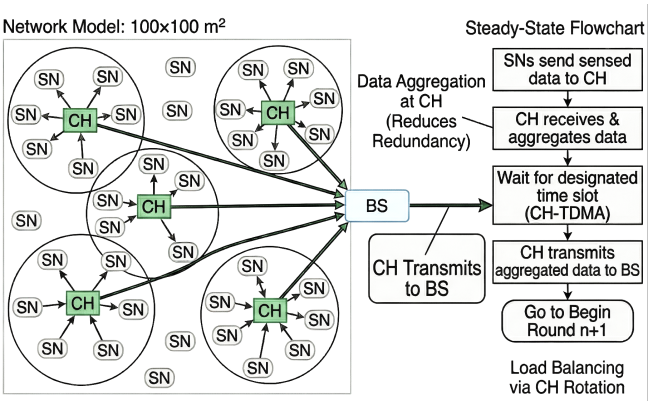


Fig. 2. Network model: N sensor nodes in k clusters, each with a CH that forwards aggregated data to the BS.

B. Energy Model

We adopt the first-order radio energy model described and analyzed in WSN surveys [19][20]. The energy consumed to transmit an l-bit data packet over distance d (see Fig. 3):

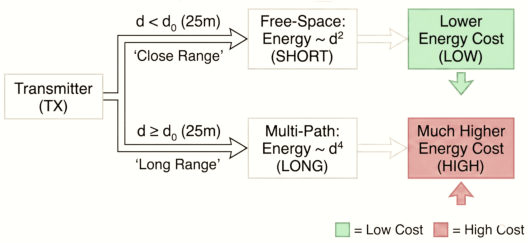


Fig. 3. Transmission cost depends on distance relative to threshold d_0.

- If $d < d_0$ (short distance, free-space model):

$$E_{Tx}(l,d) = l \cdot E_{elec} + l \cdot \varepsilon_{fs} \cdot d^2 \tag{1}$$

- If $d \geq d_0$ (long distance, multi-path fading model):

$$E_{Tx}(l,d) = l \cdot E_{elec} + l \cdot \varepsilon_{mp} \cdot d^4 \tag{2}$$

where, $d_0 = \sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$. Reception: $E_{Rx}(l) = l \cdot E_{elec}$. A CH received from N/k members, aggregates, and transmits to BS consuming more energy than a member, making CH selection the most critical factor. Table II presents the simulation parameters.

TABLE II. SIMULATION PARAMETERS

Parameter	Symbol	Value
Field / Nodes / BS	—	100×100m^2, N=100, (50,50)
Initial energy	E_0	1 J
Electronics / DA	E_{elec} / E_{DA}	50 nJ/bit / 5 nJ/bit/msg
Amplification (fs/mp)	ε_{fs} / ε_{mp}	10 pJ/bit/m^2 / 0.0013 pJ/bit/m^4
Threshold / Packet	d_0 / l	25 m / 4000 bits

C. Objective Function Formulation

The CH selection problem is formulated as a minimization problem using a composite objective function that simultaneously balances two competing criteria.

The Composite Fitness Function:

$$f_{obj} = \alpha \cdot f_1 + (1-\alpha) \cdot f_2 \tag{3}$$

where, α is a weight parameter (typically 0.5–0.7) that controls the relative priority between distance efficiency and energy protection.

Intra-Cluster Communication Distance (Minimize): f_1 represents the total Euclidean distance between all member sensor nodes and their assigned Cluster Heads:

$$f_1 = \sum_{i=1}^{k} \sum_{j \in c_i} d(SN_j, CH_i) \tag{4}$$

Rationale: Since transmission energy scales with d^2 or d^4, minimizing the physical distances between sensors and their CHs directly reduces the dominant source of energy consumption.

Residual Energy Ratio (Minimize): f_2 measures the energy quality of the elected CHs relative to the overall network:

$$f_2 = \frac{\sum_{j=1}^{N} E_{residual}(SN_j)}{\sum_{i=1}^{k} E_{residual}(CH_i)} \tag{5}$$

Rationale: A low f_2 value indicates that the selected CHs possess a large share of the network's total residual energy meaning they are "strong" nodes capable of sustaining the heavy CH workload.

IV. Interoperable Configuration Standardization

The "Interoperable" component of 2I-CSO addresses a practical yet often overlooked problem in real-world WSN deployments: how to ensure that all sensor nodes operate under identical energy parameters the very condition required for the intelligent stopping condition (Section VI-C) to be valid.

In reality, WSN implementation may consist of sensor nodes by diverse manufacturers or even from different production batches or firmware revisions. Differences in the energy consumption of radio electronics ($Eelec$), amplifiers (εfs, εmp), and data aggregation (EDA) may occur because of manufacturing tolerances even in one production batch. In the long run, some of the sensor nodes may receive new firmware while others may not, resulting in the emergence of configuration diversity – each node having individual energy parameters due to which the whole system becomes heterogeneous.

Our solution is based on standardizing the configuration process of LEACH's natural communication hierarchy: **BS → CH → SN**.

The central point of our solution is the Configuration Table stored in the Base Station. Table III consists of the standard energy values of the whole WSN.

TABLE III. BS Configuration Table: The Single Source of Truth for Network Energy Parameters

Parameter	Symbol	Standardized Value	Unit
Electronics energy	E_{elec}	50	nJ/bit
Free-space amplification	ε_{fs}	10	pJ/bit/m^2
Multi-path amplification	ε_{mp}	0.0013	pJ/bit/m^4
Data aggregation energy	E_{DA}	5	nJ/bit/message
Threshold distance	d_0	25	m
Packet size	l	4000	bits
Config version	V_{config}	v1.0	—

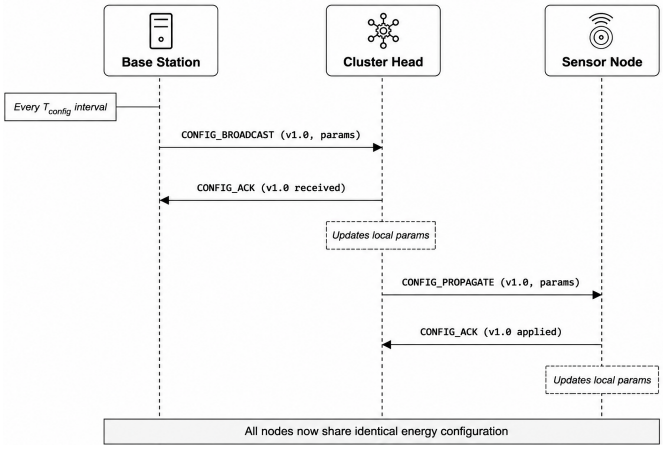


Fig. 4. Sequence diagram: scheduled interoperable configuration standardization.

Fig. 4 summarizes the scheduled configuration synchronization process, in which the BS periodically broadcasts the standardized configuration table to active CHs, each CH stores the received version and forwards it to its member SNs during intra-cluster communication, and every node uses the included version number to avoid unnecessary updates while preserving network-wide energy-parameter homogeneity.

V. Cat Swarm Optimization (CSO)

This section introduces the standard Cat Swarm Optimization algorithm as the baseline mechanism for CH selection. It first presents the biological inspiration behind CSO, then defines its main control parameters, explains the seeking and tracing search modes, and finally summarizes the standard workflow and limitations that motivate the proposed 2I-CSO enhancements.

A. Biological Inspiration

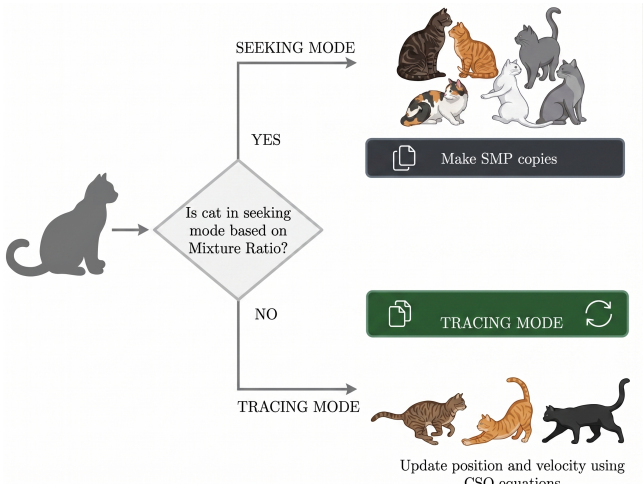


Fig. 5. CSO seeking and tracing modes: nature inspiration.

CSO, introduced in 2006 and reviewed in later CSO surveys and performance studies [12], [13], relates the cat's rest state to "Seeking mode" (local search) and the hunting state to "Tracing mode" (global search). An individual cat can be viewed as an instance of CH structure configuration such as [N2, N7]. The whole process converges towards finding the best possible configuration (gbest) for the group of cats. Benchmark studies of CSO report competitive performance against several metaheuristic algorithms on standard optimization functions [12] (see Fig. 5).

B. Parameters

The standard CSO algorithm is controlled by a small set of parameters that balance exploration and exploitation. MR determines how many cats follow the current best solution, while SMP, SRD, and CDC define how each seeking cat generates and modifies local candidate solutions. The acceleration coefficient c controls the strength of movement toward the global best configuration (see Table IV).

TABLE IV. CSO PARAMETERS

Parameter	Definition	Value
MR	Fraction in tracing mode	0.7
SMP	Number of copies generated in seeking mode	5
SRD	Mutation radius applied to selected dimensions	20%
CDC	Percentage of dimensions changed in each copy	80%
c	Tracing aggressiveness toward gbest	2.0

C. Seeking Mode: Local Exploration

A seeking cat examines nearby alternatives and moves to a superior neighbor through a structured local search:

- Create SMP copies of the current position vector.
- For each copy, randomly select CDC% of dimensions and mutate each by $\pm$SRD%.
- Evaluate the fitness f_{obj} for all copies.
- Compute selection probability: $P_i = \frac{|f_i - f_b|}{f_{max} - f_{min}}$, where f_b is the current best fitness.
- The cat is positioned where the copy chosen through the roulette wheel selection according to P_i.n

Through this process, there will be a detailed exploration of the environment around the promising solution in the manner a cat would explore its surroundings for subtle environmental changes when taking a rest.

D. Tracing Mode: Global Exploitation

The tracing cat rapidly moves towards the present global optimal solution (x_{best}):

$$v_{id} = v_{id} + r \cdot c \cdot (x_{best,d} - x_{id}) \tag{6}$$

$$x_{id} = x_{id} + v_{id} \tag{7}$$

where, $r \in [0, 1]$ is a uniformly distributed random number and c is the acceleration constant. In case of clustering of WSN nodes, the positions are converted to IDs by the rounding process: nodeID = round(x_{id}) $\bmod N$.

With MR=0.7, at each iteration 70% of cats chase (exploitation of the present optimal solution) and 30% explore (searching for alternative solutions). This dual-mode architecture prevents premature convergence to local optima while maintaining exploitation pressure toward the global optimum.

E. Standard CSO Workflow

The standard CSO procedure for WSN clustering is summarized in Algorithm 1.

Algorithm 1. Standard CSO Algorithm for WSN Clustering

Input: Sensor nodes N, number of CHs k, swarm size P, maximum iterations T, MR, SMP, SRD, CDC, c
Output: Best CH configuration $gbest$
Initialize P cats with random CH configurations $[CH_1, \ldots, CH_k]$
Evaluate each cat using $f_{obj} = \alpha f_1 + (1 - \alpha) f_2$
Set the best initial cat as $gbest$
for $t = 1$ to T **do**
  for each cat i in the swarm **do**
    Assign non-CH nodes to the nearest CH using Euclidean distance
    Generate a random value $r \in [0, 1]$
    if $r <$ MR **then**
      Apply Tracing Mode toward $gbest$
    else
      Apply Seeking Mode using SMP, SRD, and CDC
    end if
    Map updated positions to valid sensor node IDs
    Recompute f_{obj} for cat i
    if $f_{obj}(i) < f_{obj}(gbest)$ **then**
      Update $gbest \leftarrow i$
    end if
  end for
end for
return $gbest$

F. Limitations of Standard CSO

These three limitations motivate the enhancements proposed in Section VI.

1) Problem 1: Random mode switching: In standard CSO, each cat's mode is assigned randomly at every iteration according to the Mixture Ratio. A random value $r \in [0, 1]$ is generated; if $r < MR$, the cat enters tracing mode, otherwise it enters seeking mode. This mechanism is context-blind because it ignores the actual optimization state.

For example, when $gbest$ has just changed, cats should preferably trace the new optimum to exploit it. However, random assignment may still place some cats in seeking mode, causing a missed exploitation opportunity. Conversely, when $gbest$ has remained unchanged for many iterations, more cats should seek alternative regions, but standard CSO may continue assigning them to tracing mode and waste computation around a known position. Near convergence, where most cats become similar, the fixed 70% tracing and 30% seeking split can also lead to redundant movements toward the same $gbest$.

This limitation contrasts with the Emperor Penguin Optimizer, where all penguins follow a unified movement rule whose behavior changes dynamically with temperature. There is no artificial random split between modes, which makes the search behavior more adaptive.

2) Problem 2: No intelligent stopping condition: Standard CSO has no intelligent stopping condition for `chSelection()` beyond a fixed round or iteration limit. This creates two forms of computational waste. First, the optimal CH configuration may be discovered early, for example at round 20, while the algorithm continues evaluating candidates until round 2000. Second, the search may exhaust all meaningful alternatives by an early round but still continue generating redundant configurations until the fixed limit is reached.

The key insight is that existing bio-inspired WSN clustering protocols, including CSO and EPO, do not explicitly detect when CH selection has become redundant. Addressing this gap is the primary novel contribution of the proposed 2I-CSO protocol.

3) Problem 3: Configuration space explosion: Each cat in each iteration produces a CH configuration that requires cluster formation and fitness evaluation, which are among the most expensive operations in the protocol. With an ordered representation, the number of possible configurations grows

as N^k, where N is the number of sensor nodes and k is the number of selected CHs. Table V presents the combinatorial explosion of CH configuration space.

TABLE V. COMBINATORIAL EXPLOSION OF CH CONFIGURATION SPACE

Network	k	Ordered Configs	Problem
4 nodes	2	$4^2 = 16$	Manageable
10 nodes	2	$10^2 = 100$	Feasible
50 nodes	5	$50^5 \approx 3.1 \times 10^8$	Enormous
100 nodes	5	$100^5 = 10^{10}$	Intractable

Many of these configurations are redundant because the order of CHs does not change the resulting clusters. For instance, $[N_2, N_7]$ and $[N_7, N_2]$ represent the same CH set and produce identical cluster formation. Standard CSO has no mechanism to detect these repeated configurations or avoid evaluating them again.

VI. PROPOSED ENHANCEMENT: 2I-CSO

This section presents the proposed Intelligent and Interoperable Cat Swarm Optimization (2I-CSO) protocol. The objective of 2I-CSO is to improve the standard CSO search process for WSN clustering by reducing redundant CH configurations, avoiding unnecessary iterations, and adapting the swarm behavior to the current convergence state. Unlike the standard CSO, which relies on random mode assignment and a fixed stopping budget, 2I-CSO introduces a controlled search mechanism that combines configuration-space reduction, intelligent termination, adaptive mode switching, fast configuration hashing, and dynamic seeking-memory adjustment. The improvements are made to retain the optimization power of CSO, while at the same time reducing computational overhead and improving CH selection stability.

A. Overview

2I-CSO is a refined version of CSO that aims at optimizing CH selection in WSNs. Maintains the biological architecture of standard CSO (Seeking Mode for local searches and Tracing Mode for convergence) but changes the search control strategy such that it avoids redundancy and is sensitive to the state of convergence. The proposed design addresses three weaknesses of standard CSO: redundant CH configurations, fixed-budget computation, and static mode distribution.

The updated implementation introduces three main enhancement layers. First, configuration vectors are converted into canonical CH-set keys so that equivalent configurations such as $[N_2, N_7]$ and $[N_7, N_2]$ are treated as the same solution. Second, intelligent stopping is extended beyond full configuration exhaustion by adding a practical stagnation detector based on the change in global best fitness. Third, adaptive mode switching is combined with dynamic SMP decay so that 2I-CSO explores aggressively early in the search and progressively reduces expensive seeking evaluations near convergence.

B. Enhancement 1: Configuration Space Reduction

In a WSN with N active nodes and k selected CHs, a cat position represents a list of CH node IDs. Standard CSO handles this list as an ordered vector during position updates, although physically the order of CHs does not affect cluster formation. For example, $\{N_1, N_5, N_{12}\}$ and $\{N_{12}, N_1, N_5\}$ describe the same CH set and, therefore, produce the same clustering result.

The theoretical ordered search space is N^k. 2I-CSO reduces redundant tracking by converting each position into a canonical unordered configuration. In the baseline formulation, this key is obtained by sorting rounded node IDs:

$$\text{Key} = \text{tuple}\left(\text{sorted}\left(\lfloor x_i \rceil \mid x_i \in \text{position}\right)\right). \tag{8}$$

The number of unique valid configurations is then bounded by:

$$\text{Max Unique Configs} = \binom{N+k-1}{k} - N \tag{9}$$

where, the $-N$ term removes illegal configurations in which all selected CHs collapse to the same single node (see Table VI).

TABLE VI. CONFIGURATION SPACE REDUCTION EXAMPLES

Network	k	Ordered N^k	Unique Valid
4 nodes	2	16	6
10 nodes	2	100	45
100 nodes	5	10^{10}	96,560,546

C. Enhancement 2: Intelligent Stopping Condition

2I-CSO stores explored canonical configurations in a hash set $\mathcal{S}$. When the number of explored keys reaches the theoretical maximum, the CH selection engine stops because any further configuration must be redundant:

$$|\mathcal{S}| \geq \binom{N+k-1}{k} - N \Rightarrow \text{stop } \texttt{chSelection()}. \tag{10}$$

This creates a provable termination point for homogeneous search spaces and separates the expensive `chSelection()` phase from the lightweight `transmitData()` phase, which can continue until the last node dies.

In static benchmark experiments, full combinatorial exhaustion may still be difficult to reach because the valid configuration space remains large. For example, for $N = 29$ active nodes and $k = 5$ CHs, the stopping boundary is:

$$\binom{29+5-1}{5} - 29 = \binom{33}{5} - 29 = 237{,}307. \tag{11}$$

With a population of 30 cats, 500 iterations, and $SMP = 5$, the search can evaluate at most:

$$30 \times 500 \times 5 = 75{,}000 \tag{12}$$

seeking candidates. Since $75{,}000 < 237{,}307$, the complete-space stopping rule is unlikely to trigger in this benchmark setting.

To make stopping practical in benchmark and simulation settings, the updated 2I-CSO adds a delta-gbest stagnation rule:

$$\Delta f_{gbest} = |f_{gbest}^{(t-1)} - f_{gbest}^{(t)}|. \tag{13}$$

If $\Delta f_{gbest} < 10^{-6}$ in 50 successive iterations, the swarm is considered stagnant, `chSelection()` is deactivated, and the termination iteration is determined. This allows maintaining the original stopping criteria in the whole search space, with an extra option to terminate the search in large spaces.

D. Enhancement 3: Adaptive Mode Switching

Standard CSO uses random selection or tracing by setting a a fixed Mixture Ratio. In contrast, 2I-CSO selects the mode based on the condition of $gbest$:

- Active search phase: if $gbest$ changed in the previous iteration, all cats enter Tracing Mode to exploit the newly discovered promising region.
- Stable phase: if $gbest$ remains unchanged, all cats enter Seeking Mode to explore nearby alternatives and escape local stagnation.
- Initial phase: before a stable previous $gbest$ exists, cats begin with Seeking Mode to generate local diversity.

This context-aware rule keeps CSO's explicit dual-mode architecture while adopting the adaptive behavior observed in unified-motion optimizers such as EPO and SCSO [21]. Table VII presents the mode selection in Standard CSO and 2I-CSO.

TABLE VII. MODE SELECTION IN STANDARD CSO AND 2I-CSO

Optimization State	Standard CSO	2I-CSO
First iteration	Random MR split	Seeking
$gbest$ changed	Random MR split	Tracing
$gbest$ stable	Random MR split	Seeking

E. Enhancement 4: Fast Bitmask Hashing

The original canonical key based on sorting and tuple conversion is correct, but it adds overhead when called repeatedly during seeking mode. The updated implementation replaces it with an $\mathcal{O}(k)$ bitmask hash:

$$\text{Hash} = \sum_{x \in \text{position}} 2^{\lfloor x \rceil}. \tag{14}$$

For example, the CH set $\{3, 7, 12\}$ is encoded as:

$$2^3 + 2^7 + 2^{12} = 8 + 128 + 4096 = 4232. \tag{15}$$

Because addition is commutative, $[3, 7, 12]$ and $[12, 3, 7]$ generate the same key. The representation is collision-free for unique node sets because each node ID corresponds to one active bit in the integer. This avoids repeated sorting and tuple creation and improves configuration lookup efficiency.

F. Enhancement 5: Dynamic SMP Decay

The computational complexity in adaptive mode switching could be increased by switching into Seeking Mode due to the stability of $gbest$ for all cats. Given that the value of $SMP = 5$ remains fixed, this could result in numerous extra function evaluations. To avoid the above, 2I-CSO decreases the Seeking Memory Pool with respect to the iteration limit:

$$SMP_c = \max\left(2, \lfloor SMP_0 - (SMP_0 - 2)\rho \rceil\right), \quad \rho = \frac{t}{T_{max}}. \tag{16}$$

where, SMP_0 denotes the initial seeking memory pool, SMP_c the number of seeking copies currently, while ρ indicates the normalized search progress.

G. Enhancement 6: Adaptive Hybrid Mixture Ratio

The first adaptation rule had an absolute switch in that all cats were tracked whenever $gbest$ became better, and all cats searched if $gbest$ remained unchanged. Although this technique optimizes exploration or exploitation strategies, it loses the attraction aspect of the swarm when the algorithm becomes stagnated since all the cats may perform random searches without being attracted to the present gbest cat.

To preserve swarm cohesion, the optimized 2I-CSO uses an adaptive hybrid Mixture Ratio. The tracing probability is selected from the convergence state:

$$MR(t) = \begin{cases} 0.80, & \text{if } gbest \text{ changed}, \\ 0.35, & \text{if } gbest \text{ is stable} \end{cases} \tag{17}$$

In the active search phase, 80% of cats trace the newly improved leader and 20% remain in Seeking Mode for diversity. In the stagnation phase, 35% of cats maintain tracing pull toward $gbest$, while 65% perform high-pressure neighborhood search. This hybrid assignment keeps the swarm structurally aligned with the best solution while preserving the local exploration needed to escape stagnation.

H. Enhancement 7: Fitness Cache and Selective Perturbation

Because WSN fitness computation is dominated by repeated distance-energy calculations, the optimized implementation stores evaluated configurations in an exploration cache:

$$\mathcal{C} = \{\text{Hash}(\mathbf{x}) \rightarrow f(\mathbf{x})\} \tag{18}$$

Before evaluating a parent cat or a seeking copy, 2I-CSO first queries $\mathcal{C}$. If the bitmask key already exists, the cached fitness is reused in $\mathcal{O}(1)$ time; otherwise, the fitness is computed and inserted. This converts the explored-configuration structure from a pure membership set into a reusable fitness map, reducing redundant WSN evaluations during convergence plateaus.

2I-CSO also introduces selective swarm perturbation when stagnation persists for a fixed interval. The swarm is sorted by fitness, the best 50% of cats and the current $gbest$ are preserved, and only the worst 50% are relocated by randomized velocities and partial CH-node scattering. This keeps the exploitation core intact while turning weaker cats into exploration scouts that search for alternative CH basins.

I. Enhancement 8: Stagnation-Driven SMP Scaling

The linear SMP decay in Enhancement 5 reduces the seeking cost over the iteration budget. The latest implementation further adapts SMP to the stagnation depth. During active progress, SMP remains at its maximum value to preserve local-search resolution. Once a plateau is detected, the algorithm starts with a compact neighborhood and increases it only if stagnation deepens:

$$SMP_{dyn} = \max\left(2, \min\left(SMP_0, 2 + \left\lfloor \frac{s}{10} \right\rfloor\right)\right), \tag{19}$$

where, s is the consecutive stagnation counter. Thus, early plateaus are handled with a low-cost $SMP = 2$, while long stagnation gradually expands the local neighborhood up to SMP_0 to increase escape pressure.

J. Integrated Benchmarking Controls

The simulation platform provides the early termination stopping criterion as one of the benchmarks. It is off by default when there is the need to conduct an exact round-by-round comparison between CSO, EPO, and PSO, but turned on when the need is to cut down the runtime of the operations. In cases where early stopping is used, the backend stores the stop round and extends the rest of the convergence plot to ensure consistency with the sequence lengths.

The benchmark simulations will be automatically stored in the historical simulation database and can then be analyzed using the multi-run analysis interface. Thus, the proposed 2I-CSO becomes ideal for two distinct research methods – fixed-budget benchmarking and adaptive termination.

K. Performance Diagnostic

The updated implementation clarifies an important execution-time paradox. In theory, 2I-CSO should be faster than standard CSO because it can stop redundant searches. In static benchmarks, however, the full configuration boundary is often too large to reach, so the algorithm may still run for the maximum iteration budget.

Moreover, standard CSO with $MR = 0.7$ evaluates all parent cats plus only the seeking subset. With population $P = 30$ and $SMP = 5$, its approximate evaluation count per iteration is:

$$Eval_{CSO} = P + P(1 - MR)SMP = 30 + 30(0.3)(5) = 75. \quad (20)$$

When 2I-CSO enters a stable phase, all cats seek. Dynamic SMP decay controls this cost by reducing the number of copies as the iteration budget is consumed, as shown in Table VIII.

TABLE VIII. DYNAMIC SMP DECAY AND STABLE-PHASE EVALUATIONS ($P = 30$, $T_{max} = 500$).

Iter.	SMP_c	Copies	Eval.
0–83	5	150	180
84–250	4	120	150
251–416	3	90	120
417–500	2	60	90

Thus, 2I-CSO applies maximum local search when stagnation appears early, but progressively limits the seeking workload as convergence advances. The optimized version addresses the execution-time overhead through delta-gbest stopping, bitmask hashing, and dynamic SMP decay.

L. Complete 2I-CSO Algorithm

The complete 2I-CSO procedure is summarized in Algorithm 2. The algorithm begins by generating an initial population of candidate CH configurations and by creating an empty cache for previously evaluated solutions. Each candidate configuration is converted into a canonical bitmask key, which removes duplicate CH permutations and ensures that the same cluster-head set is not evaluated more than once. During each iteration, the algorithm evaluates the current population, updates the global best solution, and measures the change in best fitness to detect whether the search is still improving or has entered stagnation. If improvement is detected, the adaptive mixture ratio favors tracing mode to guide cats toward the current best solution; otherwise, it increases seeking mode and local perturbation to intensify exploration around promising regions. The process stops either when the maximum iteration budget is reached, when stagnation exceeds the allowed threshold, or when the cache confirms that all feasible unique CH configurations have been explored. The final global best configuration is then used for cluster formation and data transmission.

Algorithm 2. Optimized 2I-CSO Protocol Pseudocode

Input: Active nodes N, CH count k, population P, maximum iterations T_{max}
Initialize swarm cats with random CH positions
Initialize fitness cache $\mathcal{C} \leftarrow \emptyset$, $gbest \leftarrow \emptyset$, and stagnation counter $s \leftarrow 0$
Compute $max_configs = \binom{N+k-1}{k} - N$
for $t = 1$ to T_{max} **do**
  if $|\mathcal{C}| \geq max_configs$ or $s \geq 50$ **then**
    Deactivate `chSelection()` and record stop iteration
    break
  end if
  Evaluate all cats using cached bitmask fitness values and update $gbest$
  Compute $\Delta f_{gbest} = |f_{gbest}^{(t-1)} - f_{gbest}^{(t)}|$
  if $\Delta f_{gbest} < 10^{-6}$ **then**
    $s \leftarrow s + 1$
  else
    $s \leftarrow 0$
  end if
  Compute adaptive $MR(t)$ from the $gbest$ state
  Compute $SMP_{current}$ using iteration decay and stagnation-driven scaling
  if $s > 0$ and s is divisible by 30 **then**
    Preserve the best 50% of cats and perturb the worst 50% as exploration scouts
  end if
  for each cat c in the swarm **do**
    if $rand() < MR(t)$ **then**
      Apply Tracing Mode toward $gbest$ to maintain swarm pull
    else
      Apply Seeking Mode using $SMP_{current}$, SRD, CDC, and cached copy fitness
    end if
    Convert updated CH position to a bitmask key and store/reuse it in $\mathcal{C}$
  end for
end for
Use final $gbest$ for cluster formation and data transmission

M. Real-World WSN Operation

In static TSPLIB-style benchmarks, the active node set is fixed and the search space remains very large. In a real WSN, CH selection is executed round by round while nodes consume energy and die. Therefore, the active candidate pool N_{alive} decreases over time:

$$N_{alive}^{(r)} \leq N_{alive}^{(r-1)}. \quad (21)$$

As N_{alive} shrinks, the maximum configuration space $\binom{N_{alive}+k-1}{k} - N_{alive}$ also shrinks rapidly. For example, with $k = 5$, $N_{alive} = 15$ gives $\binom{19}{5} - 15 = 11{,}613$ unique valid configurations, while $N_{alive} = 8$ gives $\binom{12}{5} - 8 = 784$. In late network rounds, the explored set can therefore reach the complete search boundary quickly, and the intelligent stopping condition becomes increasingly effective.

The operational implication is that 2I-CSO becomes faster as the network ages. Early rounds perform deeper heuristic

search when many CH choices exist, while late rounds transition into short-circuited configuration scanning, reducing the computational energy spent on CH selection and supporting longer WSN lifetime.

VII. Simulation and Implementation

This section describes the experimental environment and benchmarking protocol used to assess the proposed 2I-CSO algorithm. The evaluation is conducted through a dedicated TSPLIB benchmarking module integrated into the WSN-Simulation application. The goal is to provide a controlled and reproducible comparison between 2I-CSO and the baseline optimization algorithms under identical execution conditions.

A. Experimental Setup

All experiments were executed on a personal computer running a 64-bit Windows 10 operating system. The hardware platform was equipped with 32 GB of RAM and an Intel Core i5 14th-generation processor with a maximum clock frequency of 5.2 GHz. The same machine was used for all TSPLIB and WSN simulation runs to ensure that runtime comparisons among CSO, 2I-CSO, EPO, PSO, and LEACH-C were performed under identical computational conditions.

B. Application Platform

To support reproducible experimentation, the WSN-Simulation application was implemented using a Next.js frontend and a Python Flask backend. The frontend provides the interface for dataset selection, parameter configuration, execution control, and visual analytics, while the Flask backend implements the optimization algorithms and returns structured benchmark results. A separate TSPLIB simulation page was added to the application in order to evaluate the proposed 2I-CSO algorithm on standardized traveling-salesman benchmark instances independently from the WSN simulation modules.

C. Rationale for Using TSPLIB in WSN Clustering Validation

Despite the fact that TSPLIB was initially created as a benchmark library for TSP problems, it could be used as a spatial optimization testing ground in this research, but not as a WSN lifetime benchmark [29]. Each TSPLIB problem presents two dimensional points for nodes together with a distance matrix for those nodes. The described characteristics could be applied to the geometric aspect of the WSN clustering, where physical deployment of sensors plays an important role as well as the dependence of communication energy from the transmission distance. As a result, TSPLIB problem could be interpreted as a static sensor deployment, where cities correspond to sensors and the chosen cities serve as candidates for CHs.

On this approach, the fitness function based on TSPLIB does not compute a TSP tour. Instead, it measures how good the clustering algorithm is at forming clusters with minimized distance clustering cost. This is useful because the typical energy models used for WSN radios are functions of distance, which can be expressed in either quadratic (d^2) or quartic (d^4) form depending on the channel conditions. Hence, clustering that minimizes average/total cluster member distance to the CH will decrease communication energy cost, make clusters more spatially compact, and allow comparison between CSO, 2I-CSO, EPO, PSO, and LEACH-C algorithms using similar node placement scenarios.

TABLE IX. Energy Depletion Benchmark for 29 Nodes, 100×100 m^2, $k = 2$

Protocol	FND	HND	LND	vs. LEACH-C (%)
CSO	608	2219	**3393**	-72.8%
2I-CSO	558	2238	3374	-75.1%
EPO	847	**2524**	3241	-62.2%
PSO	962	2220	3348	-57.0%
LEACH-C	**2239**	2246	2320	0.0%

TABLE X. FND Comparison for the Current Base Station Position.

BS Position	PSO	CSO	2I-CSO	EPO	LEACH-C
Current $(50, 50)$	962	608	558	847	**2239**

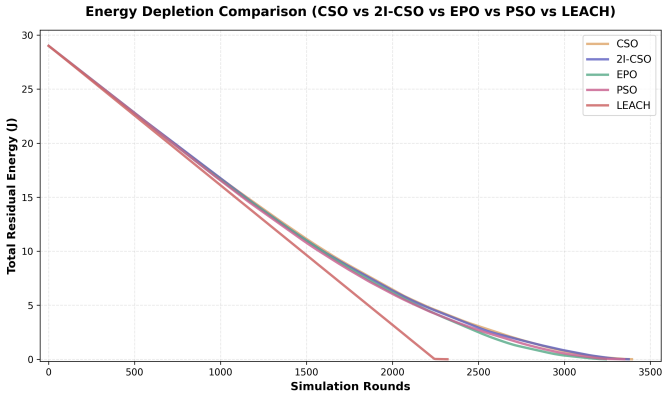
Still, TSPLIB findings need to be considered carefully. They prove that the optimizer is capable of searching large, repeatable, and distance-based CH selection spaces and of saving computation cost, convergence time, and spatial clustering fitness. Nevertheless, they are not by themselves evidence for network lifetime improvement since the TSPLIB problems lack factors such as battery depletion dynamics, traffic, CH turnover, radio interference, and node death. This is the reason why TSPLIB benchmarking is done separately as an algorithmic and geometric proof step, whereas the claims on FND, HND, LND, and lifetime are proven by the other benchmark, which follows in the next subsection.

D. WSN Energy Depletion Benchmark

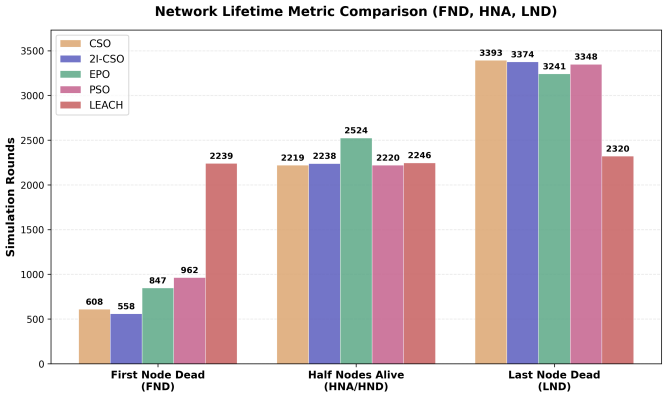
Before the TSPLIB benchmark, an additional WSN-specific energy depletion test was conducted to evaluate network lifetime behavior under a direct sensor-network scenario. The test uses 29 sensor nodes deployed in a 100×100 m^2 field, with $k = 2$ cluster heads and the Base Station located at $(50, 50)$. Three standard lifetime indicators are reported: first node death (FND), half node death (HND), and last node death (LND). The percentage column compares FND against the LEACH-C baseline (see Table IX).

Interpretation: The energy depletion results show that LEACH-C achieves the latest FND, meaning that it delays the first node failure more effectively in this deployment (see Fig. 6). However, the metaheuristic protocols maintain longer post-FND operation, with CSO and 2I-CSO producing the largest LND values. This indicates a trade-off between early stability and tail lifetime: LEACH-C is stronger when the objective is to postpone the first failure, while CSO-family methods can keep the network operating for more rounds after initial depletion begins. The 2I-CSO result remains close to CSO in LND while preserving the runtime advantages shown in the TSPLIB benchmark.

The TSPLIB webpage compares 2I-CSO against CSO, PSO, EPO, and LEACH-C in an equivalent experimentation setup (see Table X). This comparison is aimed at assessing the solution quality, convergence speed, stagnation recovery, runtime efficiency, and scalability. The experiment is carried out in five runs, 1000 iterations per run, with maximum iterations



(a) Energy depletion milestones.



(b) Network lifetime metric comparison.

Fig. 6. Grouped visualization of energy depletion and network lifetime behavior.

as the stopping criterion, similar hardware/software configurations, and similar random seeds used for each algorithm. In the case of the TSPLIB benchmark test mode, the early stagnation detection stopping rule is disabled for 2I-CSO, although the program can automatically terminate the execution when the number of consecutive stagnant runs exceeds 50. Dynamic SMP (self-multiplying population) decrement is also disabled in 2I-CSO in order to compare all algorithms on the same iteration budget without any differences in stopping rules affecting the result. This makes sure that the difference in execution time reflects only the difference in algorithmic update cost per iteration among algorithms. Population scaling factor is adaptively adjusted depending on benchmark complexity: 30 agents for small instances, 40 agents for moderate instances, 50 agents for large instances, and 60 agents for extra-large instances. This scaling preserves fairness because all algorithms use the same population size within each category (see Table XI).

The benchmark suite contains the reported TSPLIB instances distributed across three groups. The small group contains att48, bayg29, bays29, burma14, dantzig42, ulysses16, and two ulysses22 benchmark runs (see Table XII). The moderate group contains berlin52, eil51, eil76, gr96, kroa100, and krob100 (see Table XIII). The large group contains bier127, ch130, ch150, d198, eil101, and gr120 (see Table XIV). Overall, the reported experimental campaign consists of:

$$20 \times 5 \text{ protocols} \times 5 \text{ runs} = 500 \text{ executions} \qquad (22)$$

The recorded metrics include best fitness, execution time, convergence behavior, and Wilcoxon signed-rank significance results. The Wilcoxon test is applied at a level of significance of $p < 0.05$ to determine whether the observed best-fitness differences between protocols are statistically meaningful. Fig. 7 to Fig. 9 presents the TSPLIB convergence and multi-criteria radar comparison of three scales: small, moderate and large.

Table indicators have the following meanings. Best fitness corresponds to the best, or lowest, tour fitness obtained across the independent runs and measures optimization precision. Mean fitness stands for the average quality of solution, which is calculated as $\mu_f = (1/R)\sum_{i=1}^{R} f_i$ and implies expected behavior in the typical run. Standard deviation estimates the stability of the search process by measuring the variation of solution quality and is low for the most consistent optimization. Execution time shows the average CPU time per trial in seconds and measures computational efficiency. Stagnation indicators report the number and duration of non-improving intervals, while escape ratio measures the ability of the algorithm to resume improvement after stagnation.In the instance-level tables, each protocol column reports the final measured value, and the Winner column identifies the protocol with the lowest best fitness for that instance.

TABLE XI. TSPLIB BENCHMARK CONFIGURATION

Category	Instances	Population	Runs	Iterations
Small	8	30	5	1000
Moderate	6	40	5	1000
Large	6	50	5	1000

1) Final TSPLIB benchmark results: The final TSPLIB comparison evaluates five protocols: CSO, 2I-CSO, EPO, PSO, and LEACH-C. LEACH-C is included as a centralized clustering baseline because its BS-assisted CH selection is closer to the centralized configuration assumed by the proposed interoperable 2I-CSO model. For all tables, lower best-fitness values indicate better clustering cost, and execution time is reported in seconds per run. The last column reports the speedup of 2I-CSO relative to standard CSO.

Interpretation: The updated results show that 2I-CSO consistently reduces runtime relative to standard CSO across all reported TSPLIB groups, with speedups ranging from 3.2× to 56.6×. In terms of best fitness, 2I-CSO remains competitive with CSO and improves over EPO and PSO on several instances, especially in the large group. However, LEACH-C obtains the best fitness on all moderate and large instances. Therefore, the revised conclusion is that 2I-CSO offers a strong runtime–quality trade-off and improves the CSO)family computationally, while LEACH-C remains a strong centralized baseline for the TSPLIB-derived spatial clustering objective (see Table XV to Table XVII).

2) Wilcoxon signed-rank test results: The Wilcoxon signed-rank test was applied to the best-fitness values to evalu-

TABLE XII. SMALL-SCALE TSPLIB BEST-FITNESS RESULTS LOWER VALUES ARE BETTER

Instance	n	k	CSO	2I-CSO	EPO	PSO	LEACH-C	Winner
att48	48	6	**14420.8692**	14543.3432	14860.7903	14580.6611	14430.5470	CSO
bayg29	29	5	**3162.6316**	**3162.6316**	3215.8889	3185.4538	**3162.6316**	Tie
bays29	29	5	**3162.6316**	**3162.6316**	3224.5043	**3162.6316**	**3162.6316**	Tie
burma14	14	3	**10.5464**	**10.5464**	**10.5464**	**10.5464**	**10.5464**	Tie
dantzig42	42	6	288.9417	**287.1232**	296.3814	290.1005	288.7553	2I-CSO
ulysses16	16	4	**14.7519**	**14.7519**	**14.7519**	**14.7519**	14.8482	Tie
ulysses22	22	4	**19.8681**	**19.8681**	20.7483	**19.8681**	20.7483	Tie
ulysses22	22	4	**19.8681**	**19.8681**	**19.8681**	**19.8681**	20.7483	Tie

TABLE XIII. MODERATE-SCALE TSPLIB BEST-FITNESS RESULTS LOWER VALUES ARE BETTER

Instance	n	k	CSO	2I-CSO	EPO	PSO	LEACH-C	Winner
berlin52	52	7	2926.8931	2909.0199	3005.9378	3060.2780	**2905.1055**	LEACH-C
eil51	51	7	187.5013	188.8639	188.2857	190.5318	**183.1374**	LEACH-C
eil76	76	8	275.0143	275.9129	274.5026	276.2674	**265.0979**	LEACH-C
gr96	96	9	293.7943	291.7287	301.8480	312.0413	**291.0895**	LEACH-C
kroa100	100	10	12757.3741	12988.0102	13048.1546	13122.0496	**12385.5907**	LEACH-C
krob100	100	10	12857.5093	12887.3798	12860.5270	12972.6605	**12409.6372**	LEACH-C

TABLE XIV. LARGE-SCALE TSPLIB BEST-FITNESS RESULTS LOWER VALUES ARE BETTER

Instance	n	k	CSO	2I-CSO	EPO	PSO	LEACH-C	Winner
bier127	127	11	62254.5134	61863.7028	62015.0273	62191.2565	**60977.6240**	LEACH-C
ch130	130	11	4012.6637	3976.1627	4136.8095	4206.1308	**3785.5199**	LEACH-C
ch150	150	12	4543.9215	4542.9249	4500.0578	4551.2479	**4179.4261**	LEACH-C
d198	198	14	8005.7409	7735.9442	7794.9243	7749.7510	**7089.1803**	LEACH-C
eil101	101	10	331.0526	325.0657	334.2789	335.4407	**314.9213**	LEACH-C
gr120	120	10	932.9801	924.3486	960.8591	958.8654	**876.7315**	LEACH-C

TABLE XV. SMALL-SCALE TSPLIB EXECUTION TIME

Instance	CSO	2I-CSO	EPO	PSO	LEACH-C	Speedup
att48	18.514	2.954	2.383	3.299	11.555	6.3×
bayg29	10.179	2.263	1.373	1.859	10.491	4.5×
bays29	10.929	2.202	1.440	1.810	10.420	5.0×
burma14	3.234	0.997	0.607	0.772	9.708	3.2×
dantzig42	16.323	2.967	2.125	2.966	10.870	5.5×
ulysses16	4.917	1.149	0.759	1.106	10.164	4.3×
ulysses22	6.159	1.147	0.903	1.296	10.332	5.4×
ulysses22	6.812	1.365	1.018	1.412	10.202	5.0×

TABLE XVI. MODERATE-SCALE TSPLIB EXECUTION TIME

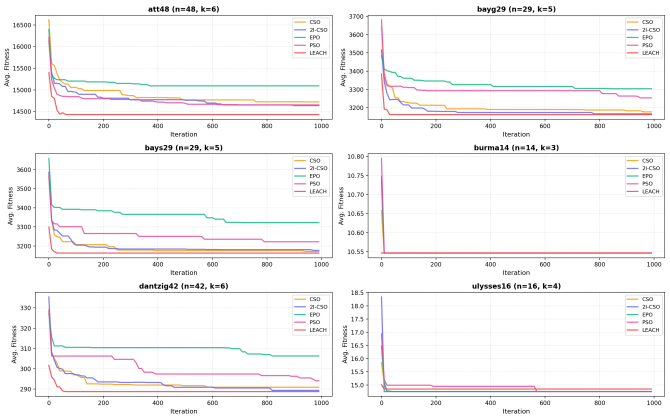
Instance	CSO	2I-CSO	EPO	PSO	LEACH-C	Speedup
berlin52	30.653	3.951	3.455	5.007	13.188	7.8×
eil51	29.213	3.693	3.305	4.711	11.981	7.9×
eil76	51.421	4.326	5.718	7.764	11.399	11.9×
gr96	73.865	4.772	7.391	10.742	12.064	15.5×
kroa100	115.141	5.183	8.766	12.184	12.226	22.2×
krob100	109.532	5.197	8.671	12.353	12.253	21.1×

TABLE XVII. LARGE-SCALE TSPLIB EXECUTION TIME

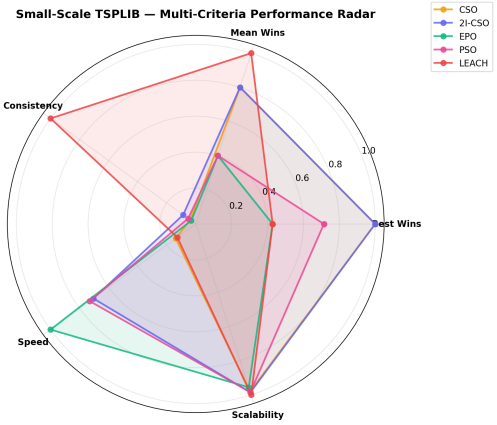
Instance	CSO	2I-CSO	EPO	PSO	LEACH-C	Speedup
bier127	144.643	7.358	14.359	20.979	29.509	19.7×
ch130	147.230	7.669	15.125	21.305	29.978	19.2×
ch150	191.401	8.517	19.370	27.815	31.345	22.5×
d198	304.949	10.689	79.614	130.372	34.434	28.5×
eil101	434.995	7.679	12.515	17.119	28.426	56.6×
gr120	142.955	7.935	14.252	19.907	28.788	18.0×

TABLE XVIII. WILCOXON SIGNED-RANK TEST RESULTS FOR TSPLIB BEST FITNESS SIGNIFICANCE THRESHOLD: $p < 0.05$

Group	Comparison	W	p-value	Significant	Outcome / Winner
Small	CSO vs 2I-CSO	2.0	0.37500	No	Statistically equivalent
Small	EPO vs 2I-CSO	0.0	0.03125	Yes	2I-CSO superior
Small	LEACH-C vs 2I-CSO	8.0	0.35938	No	Statistically equivalent
Small	PSO vs 2I-CSO	0.0	0.03125	Yes	2I-CSO superior
Moderate	2I-CSO vs CSO	9.0	0.84375	No	Statistically equivalent
Moderate	EPO vs CSO	0.0	0.03125	Yes	CSO superior
Moderate	LEACH-C vs CSO	0.0	0.03125	Yes	LEACH-C superior
Moderate	PSO vs CSO	1.0	0.06250	No	Statistically equivalent
Large	CSO vs 2I-CSO	3.0	0.15625	No	Statistically equivalent
Large	EPO vs 2I-CSO	0.0	0.03125	Yes	2I-CSO superior
Large	LEACH-C vs 2I-CSO	0.0	0.03125	Yes	LEACH-C superior
Large	PSO vs 2I-CSO	0.0	0.03125	Yes	2I-CSO superior

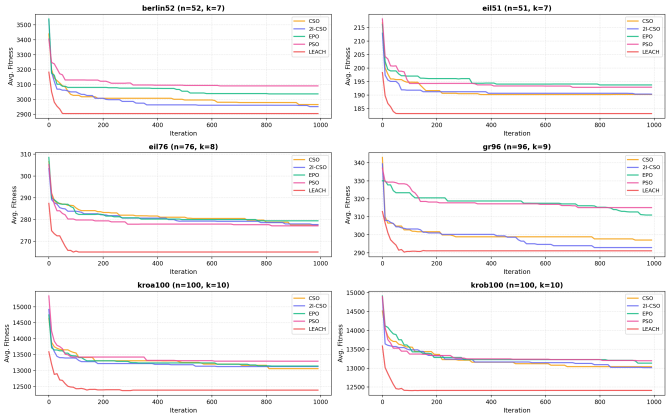


(a) Small-scale TSPLIB convergence curves: average fitness over iterations.

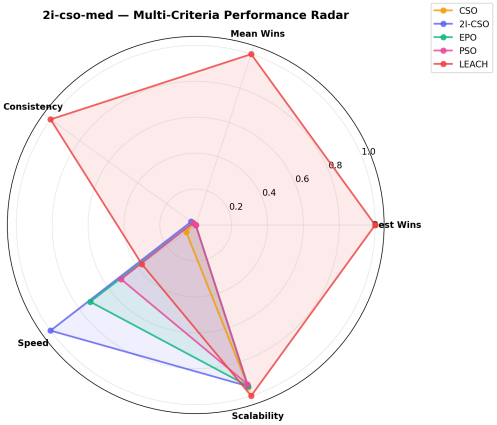


(b) Small-scale TSPLIB multi-criteria radar.

Fig. 7. Small-scale TSPLIB convergence and multi-criteria radar comparison.



(a) Moderate-scale TSPLIB convergence curves: average fitness over iterations.

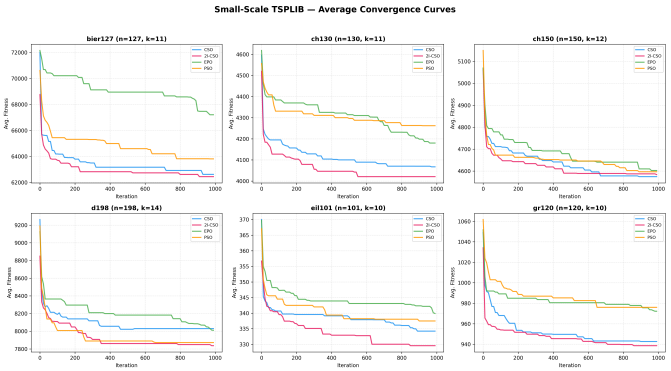


(b) Moderate-scale TSPLIB multi-criteria radar.

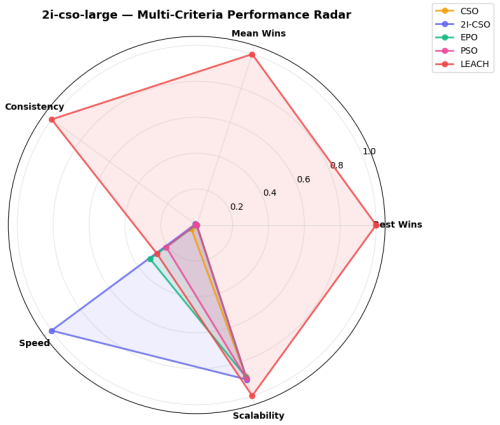
Fig. 8. Moderate-scale TSPLIB convergence and multi-criteria radar comparison.

ate whether the observed differences are statistically significant at $p < 0.05$. Table XVIII reports the final results for the small, moderate, and large TSPLIB groups. The small group uses 2I-CSO as the reference protocol. In the moderate group, the reported comparison follows the CSO-based reference used in the benchmark output. In the large group, 2I-CSO is again used as the reference protocol.

Statistical interpretation: On small instances, 2I-CSO is statistically equivalent to CSO and LEACH-C, but significantly



(a) Large-scale TSPLIB convergence curves: average fitness over iterations.



(b) Large-scale TSPLIB multi-criteria radar.

Fig. 9. Large-scale TSPLIB convergence and multi-criteria radar comparison.

better than EPO and PSO. On moderate instances, CSO and 2I-CSO are statistically equivalent, while LEACH-C is significantly better than CSO. On large instances, 2I-CSO is statistically equivalent to CSO and significantly better than EPO and PSO, but LEACH-C is significantly better than 2I-CSO. These results support a conservative interpretation: 2I-CSO improves runtime and remains competitive in fitness, but it should not be claimed to dominate LEACH-C on all TSPLIB-derived spatial benchmarks.

E. Discussion

1) Novelty of the stopping condition: The intelligent stopping criterion, as far as we know, is the first completely proven terminating approach in mathematically sound way for bio-inspired clustering of wireless sensor networks. The reason behind this is that whereas heuristic stopping criteria such as "stop if no improvement in T iterations", provide no guarantees of achieving the optimal solution, set-based monitoring guarantees that the best global solution achieved is the *true* global solution of the configuration space.

2) Synergy of enhancements: The three enhancements are not independent improvements, but form a synergistic system: Configuration Space Reduction defines the finite boundary that makes stopping possible; the Stopping Condition exploits this boundary for provable termination; and Adaptive Mode Switching accelerates convergence to reach the boundary faster. Without space reduction, the stopping threshold would be impractically large. Without adaptive switching, reaching the threshold would take longer.

3) Broader applicability: CSO has demonstrated broad applicability beyond CH selection including balanced WSN routing [18], sink node placement [25], node localization [26], underwater network partition recovery [27], target tracking [28], and NP-hard scheduling [24]. The structural nature of 2I-CSO's enhancements—operating on the search space topology rather than domain-specific fitness functions—supports their generalizability to the entire CSO family and other population-based metaheuristics.

4) Limitations:

a) Energy homogeneity requirement: The stopping condition assumes stable fitness rankings, which holds only in homogeneous networks. Heterogeneous deployments require the dynamic energy exception pathway.

b) CH candidate assumption: The formula assumes all N nodes are valid CH candidates. If certain nodes are excluded (e.g., due to location constraints), the formula must be adjusted with a reduced candidate pool $N' < N$.

c) Simulation-based validation: Results are obtained from a custom web-based simulator. Validation on a physical testbed (e.g., TelosB or MICAz motes) is necessary to confirm real-world applicability.

d) Memory scalability: When dealing with very large networks ($N > 200$, $k > 10$), Set data structure scales combinatorially and might need memory-efficient approaches (like Bloom filters, which can accept false positives).

VIII. Conclusion and Future Work

The study introduced 2I-CSO, an Intelligent and Interoperable Cat Swarm Optimization algorithm for energy-efficient Cluster Head Selection in wireless sensor network. The new approach solves many shortcomings of CSO by minimizing repetitive CH formations, incorporating a set-based intelligent termination condition, adopting a mode switching approach and unifying energy-related parameters using interoperable Base Station configuration model. The TSPLIB test cases represent a replicable spatial basis for the evaluation of searching ability, convergence characteristics, and efficiency, whereas the WSN energy exhaustion test case directly evaluates lifetime parameters such as FND, HND, and LND. The findings indicate that 2I-CSO has significantly lowered run-time relative to standard CSO and is still comparable in clustering performance,Although it appears that the lifetime study indicates that CSO-based approaches can operate for extended periods post-FND even when the delay introduced by LEACH-C in the death of the first node is greater than that achieved by 2I-CSO in the considered network. Thus, the principal advantage of the 2I-CSO approach does not consist in being superior to all existing approaches unconditionally, but

rather in a balance between runtime performance and quality based on the existence of the termination criterion and a better distinction between geometric optimizer testing and network lifetime estimation.

REFERENCES

[1] X. Liu, "A Survey on Clustering Routing Protocols in Wireless Sensor Networks," *Sensors*, vol. 12, no. 8, pp. 11113–11153, 2012.

[2] M. M. Afsar and M.-H. Tayarani-N, "Clustering in Sensor Networks: A Literature Survey," *J. Netw. Comput. Appl.*, vol. 46, pp. 198–226, 2014.

[3] S. K. Singh, P. Kumar, and J. P. Singh, "A Survey on Successors of LEACH Protocol," *IEEE Access*, vol. 5, pp. 4298–4328, 2017.

[4] S. Subha and H. Anandakumar, "Improved EPOA for WSN," *J. Electr. Eng. Technol.*, vol. 17, pp. 2499–2511, 2022.

[5] B. Singh and D. K. Lobiyal, "PSO-based CH selection for IoT-enabled WSN," *Future Gener. Comput. Syst.*, vol. 92, pp. 76–86, 2019.

[6] D. C. Hoang et al., "Fuzzy C-Means Clustering for WSNs," in *IEEE ISIE*, 2010.

[7] R. Elhabyan et al., "Pareto Optimization for WSN Clustering," *J. Netw. Comput. Appl.*, vol. 114, pp. 57–69, 2019.

[8] S. Arora and S. Singh, "Emperor Penguin Optimizer," *Knowl. Based Syst.*, vol. 164, pp. 334–345, 2019.

[9] G. Dhiman and V. Kumar, "EPO for Engineering Problems," *Knowl. Based Syst.*, vol. 159, pp. 20–50, 2018.

[10] Y. Gheraibia et al., "PeSOA," *Int. Arab J. Inf. Technol.*, vol. 16, no. 1, pp. 123–135, 2019.

[11] N. Kulkarni and P. Malathi, "PFuzzyACO for WSN," *J. Internet Technol.*, vol. 20, no. 4, pp. 1245–1254, 2019.

[12] A. M. Ahmed, T. A. Rashid, and S. A. M. Saeed, "Cat Swarm Optimization Algorithm: A Survey and Performance Evaluation," *Comput. Intell.*, vol. 36, no. 1, pp. 221–263, 2020.

[13] R. R. Ihsan et al., "A Survey on Cat Swarm Optimization Algorithm," *Asian J. Res. Comput. Sci.*, vol. 10, no. 2, pp. 22–32, 2021.

[14] Y. Kumar and P. K. Singh, "Improved Cat Swarm Optimization for Global Optimization and Clustering," *Appl. Intell.*, vol. 48, pp. 2681–2697, 2018.

[15] H. Singh and Y. Kumar, "Neighborhood Search Based Cat Swarm Optimization for Clustering," *Evol. Intell.*, vol. 15, pp. 1–23, 2020.

[16] S. Aramuthakannan et al., "Enhanced CH Selection Using Fuzzy Logic and Adaptive CSO," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 1, pp. 336–349, 2024.

[17] D. Wategaonkar and T. R. Reshmi, "CSO-Based Energy-Efficient Reliable Sectoring Scheme in WSN," *J. Inf. Sci. Eng.*, vol. 35, pp. 1249–1268, 2019.

[18] L. Kong et al., "Balanced Power Consumption via Enhanced Parallel CSO for WSN," *Sci. Program.*, vol. 2015, 913502, 2015.

[19] S. Kalburgi and M. Manimozhi, "Taylor-SCO for Energy-Efficient CH Selection with Trust Factor in WSN," *Hindawi*, 2024.

[20] M. Basheri and M. Ragab, "Quantum CSO Clustering with Intrusion Detection for IoT," *Comput. Syst. Sci. Eng.*, vol. 46, pp. 2625–2640, 2023.

[21] A. Seyyedabbasi and F. Kiani, "Sand Cat Swarm Optimization," *Eng. Comput.*, vol. 39, pp. 2627–2651, 2022.

[22] S. S. Gowda and A. Ramalingappa, "Energy Optimized CH Selection Based on M-SCSO in UWSN," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 2, pp. 304–317, 2024.

[23] S. Sindian et al., "Hybrid Dynamic Sand Cat Swarm Optimisation for WSN Resource Allocation," *IET Wirel. Sens. Syst.*, 2026.

[24] A. Bouzidi, M. E. Riffi, and M. Barkatou, "CSO for Open Shop Scheduling," *J. Ind. Manag. Optim.*, vol. 15, no. 2, pp. 733–753, 2019.

[25] V. Snasel et al., "Sink Node Placement Strategies Based on CSO," *J. Netw. Intell.*, vol. 1, no. 2, pp. 52–60, 2016.

[26] J.-S. Pan et al., "CSO-Based WSN Node Localization Algorithm," *J. Internet Technol.*, vol. 23, no. 5, pp. 949–957, 2022.

[27] M. Rao and N. K. Kamila, "CSO-Based Autonomous Recovery from Network Partitioning in UWSN," *Int. J. Syst. Assur. Eng. Manag.*, vol. 12, pp. 1125–1137, 2021.

[28] S. Tabatabaei, "CSO-Based Energy Efficient Target Tracking in WSN," *Wirel. Pers. Commun.*, vol. 137, pp. 1–24, 2024.

[29] G. Reinelt, "TSPLIB—A Traveling Salesman Problem Library," *ORSA Journal on Computing*, vol. 3, no. 4, pp. 376–384, 1991.

Emotion-Aware Serendipity Recommendation from Textual Reviews: BERT-Based Emotional Clustering

Mariam Benayad, Ahmed Zellou
SPM Research Team, ENSIAS Mohammed V University in Rabat, Rabat, Morocco

***Abstract*—This study proposes an emotion-aware serendipity recommendation framework based on textual reviews. Six BERT-based binary classifiers are trained to detect surprise, curiosity, trust, nostalgia, frustration, and enchantment from user reviews. The predicted emotion probabilities are used to build stable review-level emotional vectors. Two complementary contributions are introduced. Contribution 1 develops an emotion-aware reranking strategy that combines calibrated emotional information with relevance, novelty, and unexpectedness. Contribution 2 introduces EC-SereRank, an emotional-combination-based reranking model that preserves co-activated emotions within reviews and groups them using K-medoids into interpretable affective meta-groups. Experiments on Amazon Product Reviews show that affective information improves serendipity-oriented recommendation. The ablation study confirms that relevance and unexpectedness alone are not sufficient, while the integration of emotional compatibility improves recommendation quality. EC-SereRank further shows that emotional combinations provide an interpretable complementary signal for affective serendipity recommendation.**

***Keywords*—Recommender systems; emotion-aware recommendation; textual reviews; BERT; K-medoids; K-means**

I. Introduction

Recommender systems have become essential tools for helping users navigate large volumes of online information, products, and services [1], [2]. In e-commerce platforms, they are mainly designed to predict user preferences and generate personalized Top-K recommendation lists. Traditional recommender systems usually focus on accuracy-oriented objectives, such as relevance and ranking quality. However, highly accurate recommendations may still remain predictable, repetitive, or too similar to the user's previous interactions. This limitation can reduce the user's opportunity to discover novel and unexpected items [3], [4].

To address this issue, recent research has emphasized beyond-accuracy objectives, particularly novelty, unexpectedness, diversity, and serendipity [3], [9]. Serendipitous recommendation aims to suggest items that are not only relevant to the user, but also unexpected and potentially valuable. In this context, a good recommendation should not only match the user's known preferences, but also support discovery by introducing useful and surprising items [3], [11].

Recent studies on serendipity in recommender systems emphasize that effective recommendation should go beyond accuracy by considering novelty, unexpectedness, usefulness, and discovery value [3], [4]. Existing approaches have explored different ways to improve serendipity, including self-serendipity preference generation for cold-start scenarios [10], opinion mining from textual reviews for generating novel and pleasantly surprising recommendations [11], and multi-criteria Top-N recommendation methods [19]. In parallel, review-aware recommender systems have shown that textual reviews can improve user and item representation learning [12], [13], while sentiment-aware approaches demonstrate the usefulness of affective information extracted from user-generated content [14], [15]. However, most of these works rely either on rating patterns, sentiment polarity, or general opinion signals, and do not explicitly exploit fine-grained emotional dimensions for serendipity-oriented reranking. With the development of transformer-based language models such as BERT and DistilBERT [16], [17], and the availability of fine-grained emotion modeling resources such as GoEmotions [18], it becomes possible to build richer affective representations from textual reviews. This motivates the proposed emotion-aware framework, which integrates calibrated BERT-based emotional signals into serendipity-oriented recommendation.

Most existing serendipity-oriented approaches rely mainly on rating patterns, item popularity, user-item similarity, or structural information [10], [11]. Although these signals are useful, they do not fully exploit the rich information contained in user-generated textual reviews. Reviews provide more than explicit preference feedback; they also express subjective experiences, affective reactions, satisfaction, surprise, trust, frustration, curiosity, or emotional attachment. These affective cues can provide a deeper understanding of user preferences. Prior review-aware recommendation models have shown that textual reviews can improve user and item representation learning [12], [13], while serendipity-oriented studies suggest that richer preference signals may support discovery-oriented recommendation [3], [11].

Several studies have integrated textual reviews or sentiment analysis into recommender systems [12], [13], [14], [15]. However, sentiment-based approaches often reduce user opinions to coarse polarity categories, such as positive, negative, or neutral. Such representations are limited when the objective is to model more complex affective experiences. For example, two positive reviews may express different emotional meanings: one may reflect trust and reliability, while another may express surprise or enchantment. Similarly, a negative review may indicate frustration, disappointment, or unmet expectations. Therefore, a richer emotional representation is needed to better capture the affective structure of user reviews [18].

This study proposes an emotion-aware serendipity recommendation framework based on textual reviews. The framework extracts latent emotional signals from Amazon Product Reviews using BERT-based emotion classifiers [16], [17]. Since the original dataset does not provide explicit emotion labels, six weakly supervised binary classifiers are trained

to detect surprise, curiosity, trust, nostalgia, frustration, and enchantment from user reviews.

The proposed framework introduces two complementary contributions. The first contribution, C1, is the main quantitative component of the study. It applies K-means clustering to calibrated BERT-based emotional vectors in order to identify dominant emotional structures in the review space. These calibrated emotional vectors and cluster-based profiles are then used to construct user and item affective profiles for emotion-aware reranking based on relevance, novelty, unexpectedness, and emotional compatibility.

The second contribution, C2, extends the affective representation by modeling emotional combinations. Instead of assigning each review only to a dominant emotional profile, C2 preserves the set of simultaneously activated emotions in each review. These emotional combinations are grouped using K-medoids into interpretable affective meta-groups. This combinatorial representation allows the system to capture richer multi-emotional configurations and to integrate emotional combination compatibility into the reranking process through the proposed *EC-SereRank* model. C2 is therefore presented as a complementary representational contribution that improves the interpretability of the affective space, rather than as a replacement for C1.

The main contributions of this work are summarized as follows:

- We propose an emotion-aware serendipity recommendation framework that integrates fine-grained affective signals extracted from textual reviews into Top-K recommendation.
- We train six BERT-based binary emotion classifiers to detect surprise, curiosity, trust, nostalgia, frustration, and enchantment from user-generated reviews, and calibrate their outputs using emotion-specific thresholds.
- We propose C1, an emotion-aware reranking contribution that exploits emotional vectors together with relevance, novelty, and unexpectedness to improve serendipity-oriented recommendation. K-means clustering is also used to identify interpretable dominant emotional profiles and to support user/item affective profiling.
- We propose C2, an emotional-combination-based contribution that captures co-occurring active emotions within reviews, clusters the observed emotional combinations using K-medoids into interpretable affective meta-groups, and integrates emotional combination compatibility into the proposed EC-SereRank reranking model.
- We evaluate the proposed framework using Top-K accuracy metrics and beyond-accuracy metrics, including novelty, unexpectedness, and serendipity. The experimental results and ablation analysis show that affective information provides a useful complementary signal for serendipity-oriented recommendation.

The main quantitative contribution of this work is the C1 emotion-aware reranking model, which evaluates the impact of calibrated emotional compatibility on serendipity-oriented Top-K recommendation. C2 is introduced as a complementary representational contribution that preserves multi-emotional co-activation patterns through emotional combinations and K-medoids meta-groups.

The remainder of this study is organized as follows: Section II reviews related work on recommender systems, serendipity, textual reviews, and affective recommendation. Section III presents the proposed methodology, including dataset preprocessing, emotion annotation, BERT-based emotion extraction, emotional calibration, C1, C2, and reranking. Section IV reports the experimental results and discusses the impact of the proposed affective representations. Finally, Section V concludes the study and outlines future research directions.

II. Related Work

Recommender systems are traditionally optimized for accuracy-oriented objectives, such as relevance, precision, and ranking quality [1], [2]. However, accuracy alone may lead to repetitive and predictable recommendations, reducing the user's opportunity to discover novel and unexpected items. To overcome this limitation, recent research has increasingly focused on beyond-accuracy objectives, including novelty, diversity, unexpectedness, coverage, and serendipity [5], [6].

Serendipity is generally associated with recommendations that are relevant, unexpected, and useful to the user. Kotkov et al. [3] provide a comprehensive survey of serendipity in recommender systems and highlight the diversity of its definitions and evaluation strategies. Similarly, Ziarani and Ravanmehr [4] show that serendipity remains a complex concept because its formalization and measurement vary across studies. Earlier work by Murakami et al. [7] also emphasized the importance of evaluating serendipity through recommendation-list metrics, while Adamopoulos and Tuzhilin [8] focused specifically on the role of unexpectedness in improving user satisfaction.

Several approaches have attempted to improve serendipity through popularity debiasing, long-tail recommendation, weak-tie structures, unexpectedness modeling, or multi-objective optimization. Xu et al. [10] proposed GS2-RS, a serendipity-oriented recommendation framework that generates self-serendipity preferences from rating data in order to address cold-start and filter-bubble problems. Shrivastava et al. [11] proposed an optimized framework exploiting textual reviews and opinion mining to generate recommendations that are novel, unexpected, and relevant. Duricic et al. [9] further reviewed beyond-accuracy objectives, including diversity, serendipity, and fairness, in graph neural network-based recommender systems. These works confirm the importance of balancing relevance with discovery-oriented dimensions.

More recent studies have extended serendipity-oriented recommendation beyond classical rating-based settings. Al Jurdi et al. [20] explored the role of weak ties and network-inspired clustering in improving unexpected but relevant recommendations. Kaya and Kaleli [19] proposed a Top-N recommendation method for multi-criteria collaborative filtering and evaluated beyond-accuracy dimensions such as novelty, diversity, and serendipity. Chen et al. [21] introduced a cognitive-based knowledge learning framework that integrates

cognitive psychology and knowledge graphs to improve both accuracy and beyond-accuracy objectives, including diversity and serendipity. Yu et al. [22] addressed long-tail service recommendation using cross-view and contrastive learning, emphasizing that long-tail items can improve novelty and serendipity while reducing popularity bias. From a content-oriented perspective, Yang et al. [23] showed that recommendation content attributes, including content serendipity, can influence consumers' recommendation adoption intention.

Despite these advances, most existing serendipity-oriented approaches still rely mainly on rating data, item popularity, user-item similarity, structural signals, cognitive signals, or general opinion information. They rarely integrate fine-grained emotional dimensions extracted from textual reviews into the serendipity reranking process. This limitation motivates the proposed emotion-aware framework, which uses calibrated BERT-based affective signals to enrich serendipity-oriented recommendation beyond classical relevance, novelty, and unexpectedness components.

A. Textual Reviews and Opinion-Aware Recommendation

Textual reviews provide rich information about user preferences and item characteristics. Unlike ratings, which summarize user feedback through a numerical value, reviews contain explanations, subjective opinions, product experiences, and contextual details. Review-aware recommender systems, therefore, exploit textual reviews to improve user and item representation.

Zheng et al. [12] proposed DeepCoNN, a deep cooperative neural network that jointly learns user behaviors and item properties from review text. Chen et al. [13] introduced NARRE, an attention-based review-aware model that learns the usefulness of reviews and provides review-level explanations for recommendations. These works demonstrate that textual reviews can improve recommendation performance by providing richer user-item representations than ratings alone.

III. Methodology

A. Overview of the Proposed Framework

The proposed framework aims to enhance serendipity-oriented recommendation by exploiting latent emotional signals extracted from textual reviews. Instead of relying only on explicit ratings or conventional relevance scores, the proposed approach introduces an affective layer based on BERT-inferred emotions. These emotions are then used to construct two complementary recommendation contributions.

The overall framework is composed of four main stages. First, textual reviews are processed using BERT-based emotion classifiers in order to infer six affective dimensions: surprise, curiosity, trust, nostalgia, frustration, and enchantment. Second, the predicted emotional probabilities are used to construct review-level emotional vectors. These emotional vectors prepare the data for clustering and profile construction.

Third, two affective modeling strategies are introduced. The first contribution, C1, develops an emotion-aware reranking strategy based on review-level emotional vectors. K-means clustering is also applied to these emotional vectors in order to identify interpretable dominant emotional profiles and to support the construction of user and item affective profiles. The second contribution, C2, extends the emotional representation by preserving multiple active emotions within each review. It builds combinatorial emotional signatures and groups them using K-medoids to obtain interpretable affective meta-groups.

Finally, the affective profiles obtained from C1 and C2 are integrated into a serendipity-oriented reranking stage. The reranking function combines relevance, unexpectedness, novelty, and affective compatibility in order to generate Top-K recommendations that are both useful and discovery-oriented.

The general architecture of the proposed framework is summarized in Fig. 1.

B. Dataset and Preprocessing

The experimental study is conducted on the Amazon Product Reviews dataset obtained from Kaggle. The raw dataset initially contains 568,454 reviews and 10 attributes: *Id*, *ProductId*, *UserId*, *ProfileName*, *HelpfulnessNumerator*, *HelpfulnessDenominator*, *Score*, *Time*, *Summary*, and *Text*. Each record represents a user-item review associated with a given product.

Although the raw dataset includes several metadata and rating-related fields, the proposed preprocessing stage mainly relies on *UserId*, *ProductId*, *Text*, and *Time*. The *UserId* and *ProductId* fields are used to preserve the user-item interaction structure. The *Text* field is used as the main textual source for affective information extraction. The *Time* field is preserved for the recommendation stage in order to maintain the chronological order of user interactions.

The first preprocessing step consists of preparing the textual content of each review. Missing values in the *Text* field are replaced with empty strings, and the field is converted into textual format. Reviews with empty, very short, or non-informative textual content are removed by keeping only reviews whose length exceeds 20 characters. This filtering step ensures that the retained reviews contain enough meaningful textual information for the subsequent emotion extraction and recommendation stages.

A light text normalization step is then applied. The text is converted to lowercase, leading and trailing spaces are removed, and repeated whitespace characters are replaced with a single space. This normalization is used to make the text representation more consistent while avoiding aggressive preprocessing. In particular, no stemming, lemmatization, or heavy stop-word removal is applied, since preserving the original wording remains important for capturing contextual and affective cues in user-generated reviews.

After text cleaning, the user-item interaction structure is filtered in order to support a reliable Top-K recommendation protocol. Since users with only one or two reviews do not provide enough interaction history for constructing user-level affective profiles or for chronological evaluation, only users with at least three interactions are retained. This filtering step produced a pool of eligible users from which the experimental subset was constructed.

The final experimental subset contains 10,000 reviews. It was not obtained by uniformly sampling isolated reviews

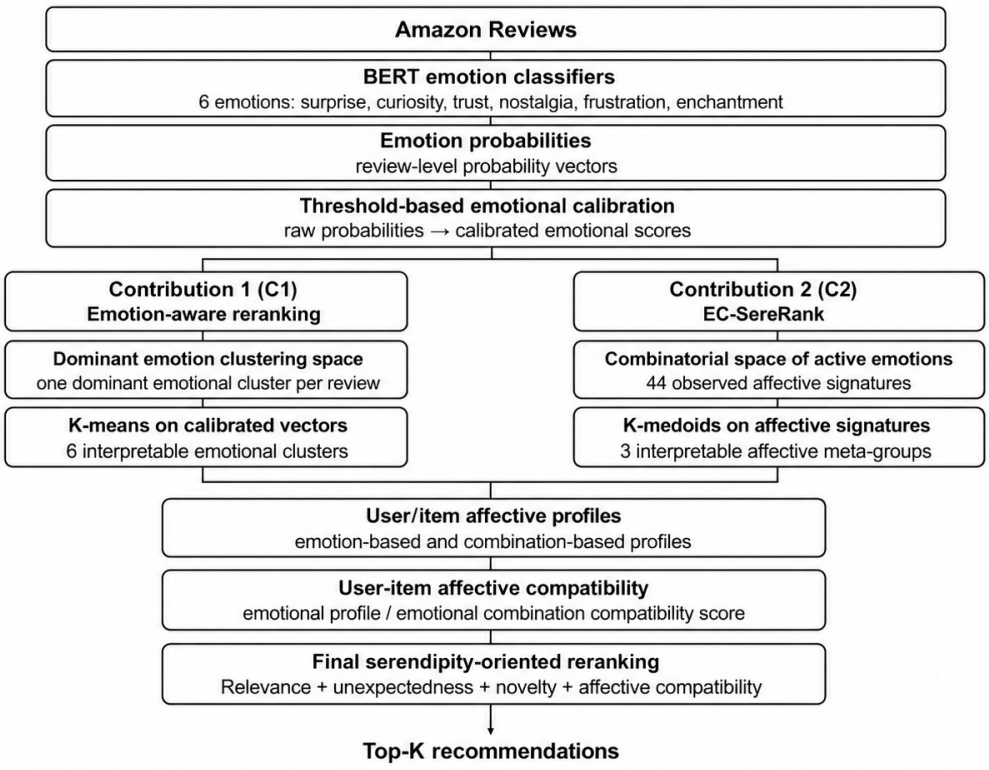


Fig. 1. General architecture of the proposed emotion-aware and combinatorial serendipity recommendation framework.

from the raw dataset. Instead, the subset was constructed from eligible users with at least three interactions, while preserving their user-item structure and temporal information. This strategy ensures that each retained user has sufficient historical interactions to support profile construction and recommendation evaluation. As shown in Table I, the resulting subset contains 1,582 users and 4,899 products. This controlled subset size also makes the BERT-based emotion extraction, clustering, user/item profile construction, and reranking stages computationally feasible while maintaining a meaningful user-item recommendation setting.

TABLE I. DATASET FILTERING SUMMARY

Step	Value
Raw reviews	568,454
Raw attributes	10
Minimum text length	> 20 characters
Minimum interactions per retained user	≥ 3
Final experimental reviews	10,000
Final users	1,582
Final products	4,899

Overall, the preprocessing pipeline transforms the raw Amazon review dataset into a structured experimental subset containing user identifiers, product identifiers, cleaned review texts, ratings, and temporal information. These prepared data are then used in the following stages of the methodology, including weakly supervised emotion annotation, BERT-based emotion extraction, affective representation learning, and Top-K recommendation evaluation.

C. Emotion Annotation Strategy

Since the original Kaggle dataset does not provide explicit emotion labels, a weak supervision strategy is adopted to generate initial affective labels from review texts. Six target emotions are considered: surprise, curiosity, trust, nostalgia, frustration, and enchantment. These emotions are selected because they cover both discovery-oriented affective states, such as surprise and curiosity, and experience-oriented affective states, such as trust, nostalgia, frustration, and enchantment.

For each emotion, a manually curated list of lexical patterns is defined using regular expressions. These patterns are designed to capture representative linguistic cues associated with each emotional dimension. For example, surprise is associated with expressions related to unexpectedness, curiosity with exploration-oriented terms, trust with reliability-related expressions, nostalgia with memory-related cues, frustration with dissatisfaction-related terms, and enchantment with delight-oriented expressions.

Before annotation, each review is normalized by converting the text to lowercase and removing redundant whitespace. The annotation function then scans the normalized text using the emotion-specific patterns. A review receives a positive weak label for an emotion when at least one corresponding lexical pattern is detected. Therefore, the minimum-hit threshold is set to one for all six emotions.

The annotation is applied independently for each emotion, making the task multi-label rather than multi-class. As a result,

the same review can be associated with several emotions when it contains cues related to multiple affective dimensions. This design is important because user reviews may naturally express mixed emotional states rather than a single isolated emotion.

After this weak annotation process, the dataset contains six binary emotion columns, one for each target emotion. These weak labels are then used to prepare emotion-specific training datasets for the BERT-based binary classifiers. One classifier is trained for each emotion using a one-vs-rest strategy. This allows the model to learn contextual affective patterns beyond the initial lexical cues used for weak labeling.

Since weak supervision may introduce noisy labels, an emotion-specific threshold calibration step is applied after BERT training. For each emotion classifier, the decision threshold is optimized on the validation set by selecting the threshold that maximizes the F1-score. The calibrated thresholds are then used to generate the final emotion vectors used in the clustering and recommendation stages.

Although this weak supervision strategy enables scalable emotion labeling without requiring a fully human-annotated corpus, it remains a limitation of the study. The generated labels should therefore be interpreted as weak affective labels rather than gold-standard human annotations. This limitation is mitigated by using contextual BERT classifiers, emotion-specific validation, and threshold calibration.

D. BERT-Based Emotion Extraction

After the weak annotation stage, emotion extraction is formulated as six independent binary classification tasks, corresponding to surprise, curiosity, trust, nostalgia, frustration, and enchantment. This formulation is consistent with the multi-label nature of user reviews, since the same review may express several emotions simultaneously. Therefore, instead of training a single multi-class model, a separate BERT-based binary classifier is trained for each emotional dimension.

For each emotion, the dataset is restricted to the review text and the corresponding binary label. The annotated data are split into training, validation, and test sets using stratified sampling in order to preserve the distribution of positive and negative classes. In the implemented pipeline, *distilbert-base-uncased* is used as the base language model. Review texts are tokenized using the BERT tokenizer, with padding and truncation applied to a maximum length of 64 tokens.

Because the emotion labels are imbalanced, a weighted training strategy is adopted. The standard cross-entropy loss is replaced with a class-weighted cross-entropy loss, giving higher importance to the minority class. This reduces the tendency of the classifier to favor the majority class, especially for less frequent emotions such as curiosity, nostalgia, and frustration.

During training, performance is monitored using accuracy and F1-score. Although a default threshold of 0.5 is commonly used for binary classification, a validation-based threshold optimization is applied after training in order to account for emotion-specific class imbalance and probability distributions. For each emotion, predicted probabilities on the validation set are evaluated over thresholds ranging from 0.10 to 0.90 with a step of 0.05. The threshold that maximizes the F1-score is selected as the optimal decision threshold; in case of equal F1-scores, the threshold with the higher accuracy is retained.

After training, the six BERT-based classifiers are applied to each review. Each classifier outputs the probability that its corresponding emotion is expressed in the review. The six predicted probabilities are then concatenated to form a six-dimensional review-level emotional probability vector. This vector represents the affective profile of the review and is used as input for the subsequent threshold-based emotional calibration stage.

E. BERT-Based Emotion Classification Results

Table II reports the final test performance of the six BERT-based emotion classifiers. The results show that the affective extraction layer provides reliable emotion signals for the downstream recommendation pipeline. Enchantment and surprise obtain the strongest F1-scores, while nostalgia, frustration, trust, and curiosity remain sufficiently robust for affective profile construction, emotional clustering, and emotion-aware reranking.

TABLE II. BERT-BASED EMOTION CLASSIFICATION RESULTS

Emotion	Threshold	Test F1
Enchantment	0.85	0.9383
Surprise	0.30	0.9207
Nostalgia	0.20	0.8418
Frustration	0.20	0.8176
Trust	0.45	0.8112
Curiosity	0.20	0.7955

F. Contribution 1: BERT-Based Emotional Clustering with K-means

The first contribution aims to structure the calibrated emotional space produced by the BERT-based emotion extraction layer. Each review is represented by a six-dimensional calibrated emotional vector composed of surprise, curiosity, trust, nostalgia, frustration, and enchantment scores. This representation provides a continuous affective description of each review.

Before applying clustering, the calibrated emotional vectors are standardized in order to ensure that all emotional dimensions contribute equally to the clustering process. K-means is then applied to the standardized calibrated vectors to identify dominant affective structures in the review space.

Given a set of calibrated emotional vectors, K-means partitions the reviews into K clusters by minimizing the within-cluster sum of squared distances:

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{r \in C_k} \|E_{cal}(r) - \mu_k\|^2 \quad (1)$$

where, C_k denotes the k-th emotional cluster, μ_k is its centroid, and $E_{cal}(r)$ represents the calibrated emotional vector of review r.

In this contribution, each review is assigned to one dominant emotional cluster. The number of clusters is set to six in order to reflect the six emotional dimensions considered

in the study. Each cluster is then interpreted according to the dominant dimension of its centroid. This produces interpretable emotional groups associated with high enchantment, high trust, high nostalgia, high frustration, high curiosity, and high surprise.

This clustering step transforms individual review-level emotion vectors into structured affective groups. Instead of using only isolated probability scores, C1 provides an interpretable emotional segmentation of the review space. These clusters are then used to support the construction of user and item emotional profiles in the recommendation pipeline.

G. C1 User and Item Emotional Profiles

After assigning each review to a dominant emotional cluster, user and item emotional profiles are constructed from the calibrated emotional vectors and the K-means cluster assignments.

For each user u, the emotional profile is obtained by aggregating the calibrated emotional vectors of the reviews associated with that user:

$$P_u^{C1} = \frac{1}{|R_u|} \sum_{r \in R_u} E_{cal}(r) \tag{2}$$

where, R_u denotes the set of reviews associated with user u.

Similarly, for each item i, the emotional profile is computed by aggregating the calibrated emotional vectors of the reviews associated with that item:

$$P_i^{C1} = \frac{1}{|R_i|} \sum_{r \in R_i} E_{cal}(r) \tag{3}$$

where, R_i denotes the set of reviews associated with item i.

In addition to these continuous emotional profiles, the distribution of K-means cluster assignments is also used to describe the dominant affective tendency of users and items. This allows the model to capture both the average emotional intensity and the dominant emotional group structure.

The affective compatibility between a user and an item is then estimated by comparing their emotional profiles:

$$AffectAlign_{C1}(u, i) = sim(P_u^{C1}, P_i^{C1}) \tag{4}$$

where, $sim(.)$ denotes a similarity function, such as cosine similarity or dot product. This affective alignment score is later integrated into the serendipity-oriented reranking stage.

H. Contribution 2: BERT-Based Combinatorial Emotional Modeling with K-Medoids

The second contribution extends the affective representation by preserving the multi-emotional nature of user reviews. While C1 assigns each review to one dominant emotional cluster, C2 represents each review through the set of emotions that are simultaneously activated. This set is referred to as an *emotional combination* and is denoted by $EC(r)$ for a given review r. This representation allows the model to capture richer affective configurations, such as trust-enchantment, surprise-curiosity, or surprise-curiosity-trust-frustration.

For each review r, an emotion e is considered active when its BERT-predicted probability exceeds its emotion-specific optimal threshold:

$$p_e(r) \geq \theta_e^* \tag{5}$$

where, $p_e(r)$ denotes the predicted probability of emotion e for review r, and θ_e^* is the optimal threshold selected during validation.

The emotional combination of a review is then formally defined as:

$$EC(r) = \{e \in E \mid p_e(r) \geq \theta_e^*\} \tag{6}$$

where, E denotes the set of the six emotions: surprise, curiosity, trust, nostalgia, frustration, and enchantment. Thus, $EC(r)$ represents the set of active emotions associated with review r.

Since six emotions are considered, the theoretical number of possible emotional combinations is:

$$2^6 = 64 \tag{7}$$

including the empty combination. In practice, only the emotional combinations observed in the dataset are retained. In the implemented pipeline, this process produces 44 observed affective signatures.

After constructing the emotional combination space, K-medoids clustering is applied to group similar emotional combinations into compact and interpretable affective meta-groups. K-medoids is selected because, unlike centroid-based clustering methods, it uses real observed emotional combinations as representative medoids. This property makes the resulting affective meta-groups easier to interpret.

Each emotional combination is represented in the six-dimensional calibrated emotional space. The dissimilarity between two emotional combinations x and y is computed using Euclidean distance:

$$d(x, y) = \sqrt{\sum_{j=1}^{6} (x_j - y_j)^2} \tag{8}$$

The K-medoids objective is to minimize the sum of distances between each emotional combination and the medoid of its assigned group:

$$\arg\min_{M} \sum_{k=1}^{K} \sum_{x \in G_k} d(x, m_k) \tag{9}$$

where, G_k denotes the k-th affective meta-group and m_k is its medoid. In this contribution, the observed emotional combinations are grouped into three interpretable affective meta-groups.

This contribution allows the framework to move beyond a single dominant emotion per review. Instead, it models emotional co-occurrences and captures affective configurations that may be particularly useful for serendipity-oriented recommendation.

I. C2 User and Item Emotional Combination Profiles

After constructing the emotional combinations and their K-medoids meta-groups, user and item profiles are built from the distribution of emotional combinations. Unlike C1, where profiles are based on averaged calibrated emotional vectors and dominant emotional clusters, C2 represents users and items through the frequency distribution of emotional combinations.

For each user u, the emotional combination profile is defined as the normalized distribution of the emotional combinations observed in the user's review history:

$$P_u^{C2} = [q_{u,1}, q_{u,2}, ..., q_{u,m}] \tag{10}$$

where, $q_{u,k}$ denotes the proportion of the k-th emotional combination in the reviews associated with user u, and m is the number of observed emotional combinations.

Similarly, for each item i, the emotional combination profile is computed as:

$$P_i^{C2} = [q_{i,1}, q_{i,2}, ..., q_{i,m}] \tag{11}$$

where, $q_{i,k}$ represents the proportion of the k-th emotional combination in the reviews associated with item i.

The compatibility between a user and an item is estimated through *emotional combination compatibility*, denoted as $ECC(u, i)$. This score is computed using cosine similarity between the user and item emotional combination profiles:

$$ECC(u, i) = \frac{P_u^{C2} \cdot P_i^{C2}}{\|P_u^{C2}\| \, \|P_i^{C2}\|} \tag{12}$$

where, P_u^{C2} and P_i^{C2} denote the user and item emotional combination profiles, respectively. The resulting score measures the degree of compatibility between the emotional combination preferences of the user and the emotional combination structure of the item.

In addition to fine-grained emotional combination profiles, the K-medoids meta-groups provide a more compact and interpretable representation of the combinatorial emotional space. Thus, C2 captures two levels of affective information: the detailed level of emotional combinations and the higher-level structure of affective meta-groups.

The resulting compatibility score $ECC(u, i)$ is later integrated into the EC-SereRank reranking model as a combinatorial affective compatibility signal

J. Serendipity-Oriented Reranking

After constructing the affective representations in C1 and C2, the final stage consists of reranking candidate items in order to promote serendipitous recommendations. The objective is to preserve relevance while introducing additional beyond-accuracy dimensions, namely unexpectedness, novelty, and affective compatibility.

For a given user u and candidate item i, the general reranking function is defined as:

$$Score(u, i) = w_r R(u, i) + w_u U(u, i) + w_n N(i) + w_a Aff(u, i) \tag{13}$$

where, $R(u, i)$ denotes relevance, $U(u, i)$ denotes unexpectedness, $N(i)$ denotes novelty, and $Aff(u, i)$ denotes the affective compatibility signal. The parameters w_r, w_u, w_n, and w_a control the contribution of each component.

In C1, the affective signal is derived from the compatibility between the continuous emotional profile of the user and the continuous emotional profile of the item. This score is denoted as emotional profile compatibility, $EPC(u, i)$:

$$Aff_{C1}(u, i) = EPC(u, i) \tag{14}$$

Thus, the C1 reranking model, denoted as ***SereEmotion***, is defined as follows:

$$SereEmotion(u, i) == w_r R(u, i) + w_u U(u, i) + w_n N(i) + w_e EPC(u, i) \tag{15}$$

In C2, the affective signal is derived from the compatibility between the emotional combination profile of the user and that of the item. This score is denoted as emotional combination compatibility, $ECC(u, i)$:

$$Aff_{C2}(u, i) = ECC(u, i) \tag{16}$$

Accordingly, the C2 reranking model, denoted as ***EC-SereRank***, is defined as follows:

$$EC{-}SereRank(u, i) = w_r R(u, i) + w_u U(u, i) + w_n N(i) + w_c ECC(u, i) \tag{17}$$

In the controlled version of ***EC-SereRank***, the emotional combination compatibility signal is injected with a small weight in order to preserve the relevance structure of the recommendation list while adding a fine-grained combinatorial affective signal:

$$EC - SereRank(u, i) = 0.95 \cdot Rel(u, i) + 0.05 \cdot ECC(u, i) \tag{18}$$

This formulation allows the model to introduce emotional combination compatibility without excessively disrupting ranking relevance. The final recommendation list is obtained by sorting candidate items according to the reranking score and selecting the top-K items.

K. Evaluation Metrics

The proposed framework is evaluated using both accuracy-oriented metrics and beyond-accuracy metrics. Accuracy-oriented metrics assess the relevance and ranking quality of the recommended items, while beyond-accuracy metrics evaluate the ability of the system to promote novelty, unexpectedness, and serendipity.

For each user u, let Rec_u^K denote the list of top-K recommended items, and let GT_u denote the set of relevant ground-truth items. Precision@K measures the proportion of relevant items among the recommended items:

$$Precision@K = \frac{1}{|U|} \sum_{u \in U} \frac{|Rec_u^K \cap GT_u|}{K} \tag{19}$$

Recall@K measures the proportion of relevant items successfully retrieved in the top-K list:

$$Recall@K = \frac{1}{|U|} \sum_{u \in U} \frac{|Rec_u^K \cap GT_u|}{|GT_u|} \tag{20}$$

HitRate@K evaluates whether at least one relevant item appears in the recommendation list:

$$HitRate@K = \frac{1}{|U|} \sum_{u \in U} \mathbb{I}(|Rec_u^K \cap GT_u| > 0) \tag{21}$$

where, $\mathbb{I}(.)$ is an indicator function.

To evaluate ranking quality, Mean Average Precision (MAP) is used. For each user, Average Precision at K is computed as:

$$AP@K(u) = \frac{1}{|GT_u|} \sum_{j=1}^{K} Precision@j(u) \cdot rel_u(j) \tag{22}$$

where, $rel_u(j) = 1$ if the item ranked at position j is relevant for user u, and 0 otherwise. MAP@K is then defined as:

$$MAP@K = \frac{1}{|U|} \sum_{u \in U} AP@K(u) \tag{23}$$

Normalized Discounted Cumulative Gain (NDCG@K) is also used to evaluate whether relevant items are ranked near the top of the recommendation list:

$$NDCG@K = \frac{1}{|U|} \sum_{u \in U} \frac{DCG@K(u)}{IDCG@K(u)} \tag{24}$$

where,

$$DCG@K(u) = \sum_{j=1}^{K} \frac{rel_u(j)}{\log_2(j+1)} \tag{25}$$

and $IDCG@K(u)$ denotes the ideal discounted cumulative gain for user u.

In addition to accuracy-oriented metrics, novelty is evaluated in order to measure the ability of the system to recommend less popular items. The novelty score of an item i is defined as:

$$Nov(i) = 1 - Pop(i) \tag{26}$$

where, $Pop(i)$ denotes the normalized popularity of item i. The novelty of a recommendation list is computed as the average novelty of its top-K items:

$$Novelty@K(u) = \frac{1}{K} \sum_{i \in Rec_u^K} Nov(i) \tag{27}$$

Unexpectedness measures the degree to which a recommended item differs from the user's expected preference space. In this work, unexpectedness is computed from the distance between user and item representations:

$$Unexp(u,i) = 1 - Sim(P_u, P_i) \tag{28}$$

where, P_u and P_i denote the user and item profiles, respectively, and $Sim(.)$ is a similarity function.

Finally, serendipity is evaluated by combining relevance with unexpectedness. The standard serendipity score is defined as:

$$Serendipity_{std}(u,i) = Rel(u,i) \cdot Unexp(u,i) \tag{29}$$

To better reflect the objective of the proposed framework, an extended serendipity metric is also considered. It incorporates relevance, unexpectedness, novelty, and affective compatibility:

$$Serendipity_{common}(u,i) = Rel(u,i) \cdot Unexp(u,i) \cdot Nov(i) \cdot Aff(u,i) \tag{30}$$

where, $Aff(u,i)$ corresponds to the affective compatibility score. In C1, this term is represented by the emotional profile compatibility $EPC(u,i)$, while in C2 it is represented by the emotional combination compatibility $ECC(u,i)$.

For a recommendation list, the serendipity score is computed as the average score over the top-K recommended items:

$$Serendipity@K(u) = \frac{1}{K} \sum_{i \in Rec_u^K} Serendipity(u,i) \tag{31}$$

L. Experimental Protocol and Ablation Study

The experimental protocol is designed to evaluate the contribution of each component of the proposed framework to Top-K recommendation performance and serendipity. The final working subset contains 10,000 reviews and is used consistently across the emotion extraction, affective profiling, and recommendation stages. Only users with sufficient interaction history are retained in order to build meaningful user profiles and support reliable evaluation.

A temporal evaluation strategy is adopted. For each user, interactions are ordered chronologically using the *Time* field. The most recent interaction is held out as the test item, while the remaining interactions are used to construct the user's historical profile. This protocol simulates a realistic recommendation scenario in which the model recommends items based on past user behavior and is evaluated on a future interaction.

The candidate items are reranked according to different scoring functions. To analyze the impact of each component, an ablation study is conducted by progressively adding unexpectedness, novelty, and affective compatibility to the relevance-based baseline. The evaluated variants are defined as follows:

- Relevance-only: uses only the relevance score $Rel(u, i)$.
- Relevance + Unexpectedness: combines relevance with unexpectedness.
- Relevance + Unexpectedness + Emotion: integrates the emotional compatibility signal from C1.
- Relevance + Novelty + Unexpectedness + Emotion: combines all C1 components.
- EC-SereRank: integrates emotional combination compatibility from C2 as a controlled complementary signal to relevance.

The purpose of this ablation study is to determine whether the proposed affective components improve recommendation quality beyond a standard relevance-driven ranking. In particular, the comparison between the relevance-only baseline and the emotion-aware variants makes it possible to quantify the added value of emotional modeling. Similarly, the comparison between C1-based emotional compatibility and C2-based emotional combination compatibility allows us to evaluate whether modeling active emotional combinations provides an additional benefit for serendipity-oriented recommendation.

All models are evaluated at different cutoff values, namely $K = 10$, $K = 20$, and $K = 50$. The evaluation includes both accuracy-oriented metrics, such as Precision@K, Recall@K, HitRate@K, MAP@K, and NDCG@K, and beyond-accuracy metrics, such as Novelty@K, Unexpectedness@K, and Serendipity@K. This dual evaluation makes it possible to assess whether the proposed models improve discovery-oriented recommendation while preserving acceptable ranking quality. The results of this ablation protocol are reported and discussed in the following section.

IV. Results and Discussion

This section presents the main experimental results of the proposed emotion-aware serendipity recommendation framework. The analysis focuses on three aspects: the performance of the BERT-based emotion classifiers, the interpretability of the affective structures obtained in C1 and C2, and the impact of the proposed affective signals on Top-K recommendation and serendipity.

A. Performance of BERT-Based Emotion Classification

The first step evaluates the quality of the BERT-based emotion extraction layer. Six independent binary classifiers were trained for surprise, curiosity, trust, nostalgia, frustration, and enchantment. Table III summarizes the final test F1-scores and the optimal thresholds selected on the validation set.

TABLE III. BERT-Based Emotion Classification Results

Emotion	Optimal Threshold	Test F1-Score
Surprise	0.30	0.9207
Curiosity	0.20	0.7955
Trust	0.45	0.8112
Nostalgia	0.20	0.8418
Frustration	0.20	0.8176
Enchantment	0.85	0.9383

The results show that the emotional layer is globally reliable. Enchantment and surprise obtain the strongest performance, while nostalgia, frustration, and trust also provide robust results. Curiosity is more difficult to detect, but remains exploitable for the recommendation pipeline. These results confirm that the extracted emotional probabilities can be used as meaningful affective signals.

B. Confusion Matrices of the Six BERT-Based Emotion Classifiers on the Test Set

To further validate the reliability of the BERT-based emotion classifiers, the confusion matrix components were visualized using two complementary figures. Fig. 2 presents the correct classifications, namely true negatives (TN) and true positives (TP), for each emotion. The results show that the models correctly identify a large number of non-active emotion instances, while also detecting positive emotion cases across the six affective dimensions. In particular, enchantment obtains the highest true positive count, which confirms the strong detection capacity already observed through its F1-score.

Fig. 3 focuses on the classification errors by reporting false positives (FP) and false negatives (FN). This representation makes it easier to identify the most challenging emotions. Trust presents the highest number of false negatives and false positives, indicating that it remains the most difficult emotion to classify. Curiosity also appears more challenging than surprise and nostalgia, but its error level remains acceptable for downstream affective recommendation. Overall, these two figures confirm that the proposed emotion classifiers do not simply predict the majority class, but provide exploitable emotional signals for the calibration, clustering, and serendipity-oriented reranking stages.

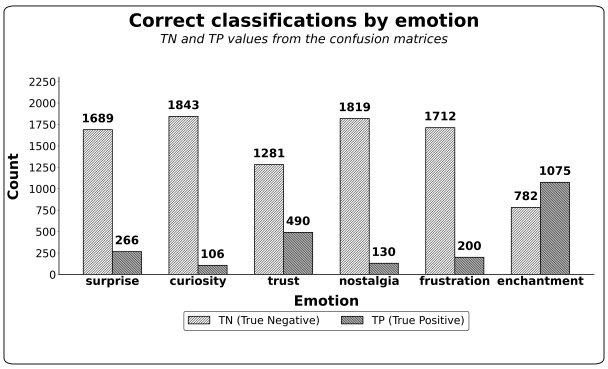


Fig. 2. Correct classification components of the six BERT-based emotion classifiers.

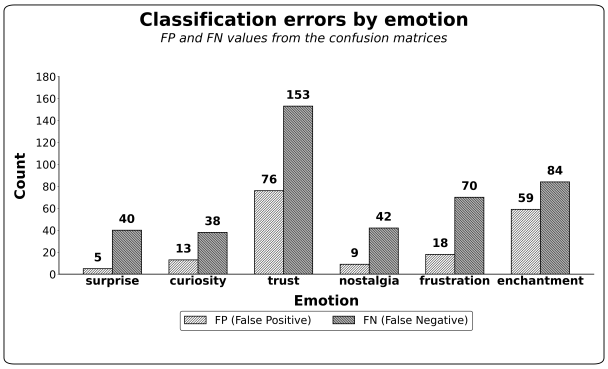


Fig. 3. Classification error components of the six BERT-based emotion classifiers.

C. Literature Positioning of C1 Against Existing Systems

Table IV positions C1 with respect to existing Top-10 recommendation systems reported in the literature. The values of the existing systems correspond to the average results over the six datasets reported in the comparative study. However, this table is provided only as a literature positioning reference. The compared systems were evaluated on different datasets and under different experimental protocols, whereas C1 is evaluated on the Amazon Fine Food Reviews subset used in this study. Therefore, the results should not be interpreted as a direct empirical comparison.

TABLE IV. LITERATURE POSITIONING OF C1 AGAINST EXISTING TOP-10 SYSTEMS.

System	Prec.	Recall	Ser.	nDCG
Matrix factorization based recommendations	0.340	0.322	0.021	–
Hybrid recommendations	0.410	0.383	0.027	0.335
Sankar et al.	0.415	0.408	0.036	0.373
FBPRR	0.422	0.445	0.030	0.382
C1: Emotion-aware reranking	0.061	**0.611**	0.034	**0.449**

The results show that C1 follows a different performance profile from the systems reported in the literature. In terms of Precision@10, C1 obtains a lower value than the compared systems. This difference should be interpreted carefully, since C1 is evaluated under a leave-one-out Top-K protocol in which each user has only one held-out relevant item. Under this setting, Precision@10 is bounded by 0.10 and corresponds to Recall@10 divided by 10. Despite this lower precision value, C1 achieves a high Recall@10 of 0.611, indicating that the proposed model is able to recover the held-out relevant item for approximately 61% of the evaluated users. C1 also obtains a competitive nDCG@10 value of 0.449, suggesting that the relevant item is often ranked in a favorable position within the Top-10 list. Regarding serendipity, C1 reaches a value of 0.034, which is competitive with the values reported by the compared systems. Nevertheless, because serendipity definitions and evaluation protocols differ across studies, this result should be interpreted as a positioning indicator rather than as a direct superiority claim. Overall, the table shows that C1 is not optimized for precision alone, but provides a strong recall-oriented and serendipity-oriented recommendation profile.

D. Ablation Analysis of C1: Added Value of the Emotional Component

To evaluate the contribution of the emotional component in C1, an ablation analysis was conducted by progressively integrating relevance, unexpectedness, emotion, and novelty into the reranking function. Table V reports the serendipity-oriented results at $K = 10$.

TABLE V. ABLATION RESULTS OF C1 AT K=10

C1 Variant	Ser_std	Ser_full
Relevance only	0.000488	0.012546
Relevance + Unexpectedness	0.000715	0.012714
Relevance + Unexpectedness + Emotion	0.002509	0.024797
Relevance + Novelty + Unexpectedness + Emotion	**0.005631**	**0.033690**

While Fig. 4 and Fig. 5 provide a graphical representation of the same ablation behavior for the standard and full serendipity metrics, respectively.

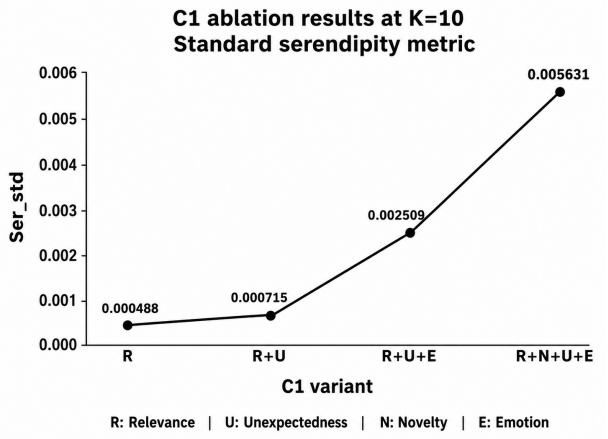


Fig. 4. Ablation results of C1 at $K = 10$ according to the standard serendipity metric.

The ablation results show that relevance alone is not sufficient to generate strong serendipitous recommendations. Adding unexpectedness slightly improves the serendipity score, but the gain remains limited. In contrast, the integration of the emotional component produces a clear improvement: the model combining relevance, unexpectedness, and emotion achieves a substantially higher serendipity score than the relevance–unexpectedness variant. This confirms that emotional alignment provides an effective complementary signal for serendipity-oriented reranking. The best performance is obtained when novelty is added together with relevance, unexpectedness, and emotion, which indicates that C1 benefits most from the joint integration of affective, novel, and unexpected recommendation signals.

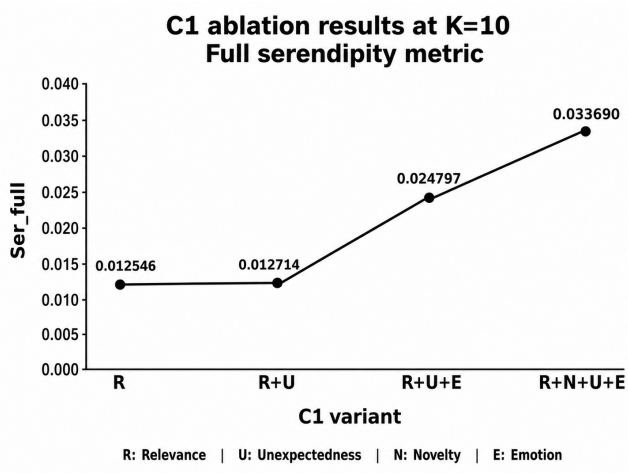


Fig. 5. Ablation results of C1 at $K = 10$ according to the full serendipity metric.

E. Analysis of C2: Emotional Combination Modeling with K-Medoids

C2 extends C1 by moving from dominant emotional clustering to fine-grained emotional combination modeling. While C1 assigns each review to one dominant emotional cluster, C2 preserves the multi-emotional nature of reviews by representing each review through its active emotional combination, denoted as $EC(r)$. With six calibrated emotions, the theoretical number of possible combinations is $2^6 = 64$. In practice, 44 emotional combinations were observed in the dataset.

K-medoids clustering was then applied to these observed emotional combinations in order to obtain compact and interpretable affective meta-groups. The final retained solution produced three meta-groups, summarized in Table VI.

TABLE VI. K-MEDOIDS AFFECTIVE META-GROUPS IN C2

Meta-Group	Reviews	Percentage
Positive affective core	5528	55.28%
Surprise discovery	2361	23.61%
Trust exploratory tension	2111	21.11%

These results show that C2 captures richer affective structures than C1. Instead of reducing each review to a single dominant emotional cluster, C2 keeps the interaction between multiple activated emotions. The first meta-group represents a positive affective pole mainly associated with enchantment, the second captures a discovery-oriented surprise profile, and the third reflects a trust-based exploratory tension. Therefore, C2 provides an additional level of affective interpretation by modeling emotional configurations rather than isolated dominant emotions.

F. Added Value of C2 Compared with C1

The added value of C2 compared with C1 lies mainly in the way affective information is represented and exploited in the recommendation process. C1 provides a clear and interpretable segmentation of the emotional space by assigning each review to a dominant affective cluster. This representation is useful because it identifies the main emotional tendency of each review. However, it may simplify reviews that contain several simultaneous emotional signals.

C2 addresses this limitation by preserving emotional co-activation patterns through emotional combinations. Instead of reducing a review to a single dominant emotional cluster, C2 represents each review by the set of active emotions detected by the calibrated BERT-based classifiers. This set is formalized as an emotional combination, denoted by $EC(r)$. Thus, a review can be represented not only by a single emotion such as *trust* or *surprise*, but also by richer affective signatures such as *trust-enchantment*, *surprise-curiosity*, or *surprise-curiosity-trust-frustration*.

TABLE VII. CONCEPTUAL ADDED VALUE OF C2 COMPARED WITH C1

Aspect	**C1 / Emotion-Aware Reranking**	**C2 / EC-SereRank**
Affective unit	Dominant emotional cluster	Emotional combination $EC(r)$
Representation	Single dominant affective profile	Multi-emotional signature
Emotional modeling	Main emotional tendency	Co-activated emotional patterns
Clustering method	K-means	K-medoids
Interpretation	Six emotional groups	Affective meta-groups
Recommendation signal	Emotion alignment	Emotional combination compatibility
Main value	Emotion-aware reranking and segmentation	Fine-grained affective compatibility

Based on this representation, EC-SereRank introduces an emotional combination compatibility component into the reranking process. This component measures the compatibility between the user and the item at the level of observed emotional configurations. Therefore, EC-SereRank captures finer affective relationships than a representation based only on individual emotions. This is particularly important for serendipity-oriented recommendation, where the objective is not only to recommend relevant items, but also to promote unexpected and emotionally meaningful discoveries.

Table VII summarizes the conceptual added value of C2 compared with C1. While C1 mainly provides an interpretable

TABLE VIII. EMPIRICAL GAINS OF EC-SERERANK OVER THE RELEVANCE-ONLY BASELINE IN C2

K	Precision Gain	Recall Gain	MAP Gain	NDCG Gain	Serendipity Common Gain
10	+8.60%	+8.60%	+13.97%	+12.14%	+12.03%
20	+11.78%	+11.78%	+14.40%	+13.37%	+7.06%
50	+14.43%	+14.42%	+14.61%	+14.27%	+6.01%

emotion-aware segmentation, C2 enriches the framework by introducing a fine-grained affective compatibility layer through EC-SereRank.

The empirical results further confirm the usefulness of the combinatorial signal inside the C2 framework. Compared with the relevance-only baseline, EC-SereRank improves the main ranking metrics and the proposed common serendipity score at different values of K. These gains show that emotional combinations provide a useful complementary signal when they are injected into the reranking function with a controlled weight.

As shown in Table VIII, EC-SereRank consistently improves the relevance-only baseline in Precision, Recall, MAP, NDCG, and Serendipity Common. At $K = 10$, the model improves Precision and Recall by 8.60%, MAP by 13.97%, NDCG by 12.14%, and the common serendipity metric by 12.03%. Similar improvements are observed at $K = 20$ and $K = 50$, which confirms that emotional combinations contribute positively to the recommendation process.

Nevertheless, C2 also has some limitations. First, its absolute Top-K performance remains lower than the best C1 model. This suggests that the combinatorial representation may lose part of the fine-grained continuous affective information captured by C1. Second, the variants that include novelty, unexpectedness, and meta-group information do not always improve the results compared with EC-SereRank. This indicates that additional C2 components may require more careful weighting. Third, the standard serendipity score slightly decreases compared with the relevance-only baseline, while the common serendipity score improves. This means that C2 is especially beneficial when serendipity is evaluated through a definition that includes affective compatibility.

Overall, C1 remains the strongest contribution from a purely quantitative recommendation perspective, while C2 provides a richer and more interpretable affective modeling strategy. The main value of EC-SereRank is not that it replaces C1, but that it introduces a more detailed emotional representation able to capture co-occurring affective signals and use them as complementary evidence for serendipity-oriented recommendation.

G. Discussion

Overall, the results support the main hypothesis of this work: affective information extracted from textual reviews can improve serendipity-oriented recommendation. C1 shows that calibrated BERT-based emotional vectors can be effectively exploited to build an emotion-aware reranking strategy. The results demonstrate that integrating emotional signals into the recommendation process improves both ranking quality and serendipity-oriented metrics compared with relevance and unexpectedness alone.

The ablation study confirms that relevance and unexpectedness are not sufficient to produce strong serendipitous recommendations. Adding unexpectedness alone leads to only a limited improvement, whereas the integration of affective information produces a more substantial gain. The best results are obtained when relevance, novelty, unexpectedness, and emotional compatibility are jointly considered. This confirms that emotion acts as a complementary signal that helps identify recommendations that are not only relevant, but also more meaningful from an affective and discovery-oriented perspective.

C2 extends this contribution by introducing a richer representation of affective information. While C1 mainly relies on dominant emotional tendencies, C2 preserves emotional co-activation patterns through emotional combinations. This allows EC-SereRank to model fine-grained affective compatibility between users and items. The empirical gains obtained by EC-SereRank over the relevance-only baseline show that emotional combinations provide useful additional information within the C2 framework, especially when evaluated through the proposed common serendipity metric.

However, the results also show that C2 should not be interpreted as a complete replacement for C1. C1 remains stronger in terms of absolute Top-K performance, while C2 provides a more expressive and interpretable affective modeling layer. Therefore, the two contributions are complementary: C1 provides the strongest quantitative emotion-aware reranking model, whereas C2 introduces EC-SereRank as an advanced combinatorial extension that improves the granularity and explainability of affective recommendation.

These findings suggest that serendipity-oriented recommendation can benefit from moving beyond classical relevance-based ranking. By incorporating calibrated emotional signals and emotional combinations, the proposed framework offers a more nuanced way to connect users and items through affective compatibility. Nevertheless, the evaluation remains offline, and future work should validate whether these affective and serendipitous recommendations are also perceived as useful, surprising, and satisfying by real users.

Therefore, the proposed framework contributes in two complementary ways. First, it introduces an emotion-aware reranking strategy based on calibrated BERT outputs. Second, it proposes a combinatorial emotional representation that captures co-occurring affective signals and exploits them through EC-SereRank. Together, these two contributions show that textual emotions can be transformed into actionable recommendation signals for improving serendipity.

V. Conclusion and Future Work

This study proposed an emotion-aware serendipity recommendation framework based on textual reviews. The main objective was to move beyond relevance-driven recommendation by exploiting affective information extracted from user-generated reviews. To this end, weak emotional annotation and BERT-based binary classifiers were used to detect six emotional dimensions: surprise, curiosity, trust, nostalgia, frustration, and enchantment. A threshold-based calibration step was then applied to obtain more stable and comparable emotional representations.

Two complementary contributions were introduced. C1 developed an emotion-aware reranking strategy based on calibrated emotional vectors, while C2 extended this representation through emotional combinations and EC-SereRank. The first contribution demonstrated the effectiveness of direct emotional integration in the reranking process, whereas the second provided a richer and more interpretable affective representation by preserving emotional co-activation patterns.

The experimental results confirm that affective information can improve serendipity-oriented recommendation. The ablation study showed that relevance and unexpectedness alone are not sufficient to achieve strong serendipitous recommendations. The best results were obtained when relevance, novelty, unexpectedness, and emotional compatibility were jointly considered. In addition, EC-SereRank showed that fine-grained emotional combinations can provide a useful complementary signal within the C2 framework.

Despite these promising results, some limitations remain. The evaluation was conducted in an offline setting and on a single dataset, which limits the ability to fully assess real user perception and cross-domain generalization. Moreover, the performance of the framework depends on the quality of the emotion classifiers and on the calibration thresholds.

Future work will focus on validating the proposed framework on additional datasets and domains, integrating user-centered or online evaluation, and improving the weighting strategy used in the reranking process. Another promising direction is to combine emotional combinations with explainable recommendation mechanisms in order to provide users with clearer reasons for serendipitous recommendations.

References

[1] F. Ricci, L. Rokach, and B. Shapira, Eds., *Recommender Systems Handbook*, 2nd ed. New York, NY, USA: Springer, 2015.

[2] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[3] D. Kotkov, S. Wang, and J. Veijalainen, "A survey of serendipity in recommender systems," *Knowledge-Based Systems*, vol. 111, pp. 180–192, 2016.

[4] R. J. Ziarani and R. Ravanmehr, "Serendipity in recommender systems: A systematic literature review," *Journal of Computer Science and Technology*, vol. 36, no. 2, pp. 375–396, 2021, doi: 10.1007/s11390-020-0135-9.

[5] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: Evaluating recommender systems by coverage and serendipity," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, 2010, pp. 257–260.

[6] M. Kaminskas and D. Bridge, "Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 7, no. 1, pp. 1–42, 2016.

[7] T. Murakami, K. Mori, and R. Orihara, "Metrics for evaluating the serendipity of recommendation lists," in *New Frontiers in Artificial Intelligence*, Lecture Notes in Computer Science, vol. 4914, Springer, 2008, pp. 40–46.

[8] P. Adamopoulos and A. Tuzhilin, "On unexpectedness in recommender systems: Or how to better expect the unexpected," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 4, pp. 1–32, 2014.

[9] T. Duricic, D. Kowald, E. Lacic, and E. Lex, "Beyond-accuracy: A review on diversity, serendipity and fairness in recommender systems based on graph neural networks," *Frontiers in Big Data*, vol. 6, p. 1251072, 2023, doi: 10.3389/fdata.2023.1251072.

[10] Y. Xu, Y. Yang, and E. Wang, "Generating self-serendipity preference in recommender systems for addressing cold start problems," *arXiv preprint arXiv:2204.12651*, 2022.

[11] R. Shrivastava, D. S. Sisodia, N. K. Nagwani, and U. Roy, "An optimized recommendation framework exploiting textual review based opinion mining for generating pleasantly surprising, novel yet relevant recommendations," *Pattern Recognition Letters*, vol. 159, pp. 91–99, 2022.

[12] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 425–434.

[13] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," in *Proceedings of the 2018 World Wide Web Conference Companion*, 2018, pp. 1583–1592.

[14] C. N. Dang, M. N. Moreno-Garcia, and F. De la Prieta, "An approach to integrating sentiment analysis into recommender systems," *Sensors*, vol. 21, no. 16, p. 5666, 2021.

[15] S. M. Al-Ghuribi and S. A. M. Noah, "A comprehensive overview of recommender system and sentiment analysis," *arXiv preprint arXiv:2109.08794*, 2021.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[18] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[19] T. Kaya and C. Kaleli, "A novel top-n recommendation method for multi-criteria collaborative filtering," *Expert Systems with Applications*, vol. 198, p. 116695, 2022, doi: 10.1016/j.eswa.2022.116695.

[20] W. Al Jurdi, J. Bou Abdo, J. Demerjian, and A. Makhoul, "Exploring the power of weak ties on serendipity in recommender systems," in *Complex Networks 2023*, Studies in Computational Intelligence, vol. 1144, Springer, 2024, pp. 205–216, doi: 10.1007/978-3-031-53503-1_17.

[21] X. Chen, Q. Liang, Y. Chen, P. Wang, H. Yu, and X. Luo, "Cognitive-based knowledge learning framework for recommendation," *Knowledge-Based Systems*, vol. 287, p. 111446, 2024, doi: 10.1016/j.knosys.2024.111446.

[22] D. Yu, T. Yu, D. Wang, and S. Wang, "Long tail service recommendation based on cross-view and contrastive learning," *Expert Systems with Applications*, vol. 238, p. 121957, 2024, doi: 10.1016/j.eswa.2023.121957.

[23] T. Yang, F. Yang, and J. Men, "Recommendation content matters! Exploring the impact of the recommendation content on consumer decisions from the means-end chain perspective," *International Journal of Information Management*, vol. 68, p. 102589, 2023, doi: 10.1016/j.ijinfomgt.2022.102589.

QuantumGuard: A Post-Quantum Resilient Deception-Driven Framework for Proactive Threat Hunting and Cognitive Honeynet Orchestration in 6G-Enabled Cyber-Physical Systems

Daifallah Zaid Alotaibe
Software Engineering Department, College of Computer Science and Engineering
University of Hafr Al-Batin, Saudi Arabia

***Abstract*—The convergence of 6G ultra-reliable low-latency communications, massive cyber-physical actuation, and the looming threat of cryptographically relevant quantum computers exposes a fundamentally new attack surface that is poorly addressed by reactive intrusion-detection paradigms. This study presents QuantumGuard, a proactive cybersecurity framework that inverts the conventional defender posture by integrating cognitive honeynets, post-quantum-secured intelligence channels, and reinforcement-learning-driven deception adaptation across 6G-enabled cyber-physical environments. The principal contribution is the architectural integration of four previously disjoint capabilities—a cognitive honeynet orchestrator, a reinforcement-learning deception policy engine operating under partial observability, a transformer-based MITRE ATT&CK attribution network, and a post-quantum federated intelligence bus secured with CRYSTALS-Kyber and CRYSTALS-Dilithium—into a single closed-loop control architecture for 6G cyber-physical systems. To assess feasibility, we report a preliminary evaluation on the CICAPT-IIoT-2024 and Edge-IIoTset benchmark datasets, complemented by a 320-endpoint 6G slice testbed configured, as described in Section IV. Initial results indicate an attacker engagement-retention rate of approximately 96.42 per cent, MITRE ATT&CK technique-attribution accuracy of approximately 91.8 per cent, a 78.6 per cent reduction in median attacker dwell time relative to passive honeypot baselines, and a deception-induced production-traffic overhead of 1.7 per cent. The PQ-FIB sustains a 14.2 ms median post-quantum handshake latency at slice scale. We position these numbers as preliminary evidence of operational viability under the evaluated threat model rather than as fully characterized performance bounds; a follow-up empirical study, planned for a separate publication, will extend the evaluation to deception-aware adversaries and sustained-attack stress conditions.**

***Keywords*—Cyber deception; cognitive honeynet; post-quantum cryptography; reinforcement learning; 6G security; cyber-physical systems; threat attribution**

I. Introduction

The deployment of 6G mobile networks, with their promise of sub-millisecond latency, terabit-per-second throughput, and pervasive integration of cyber-physical systems (CPS), is reshaping critical infrastructure ranging from autonomous transportation and smart grids to remote surgery and industrial automation [1], [2], [3]. At the same time, the advent of cryptographically relevant quantum computers threatens to undermine the public-key cryptographic primitives that currently protect this infrastructure [4], [5], [6]. Sophisticated adversaries are already adopting harvest-now-decrypt-later strategies, exfiltrating ciphertext today in anticipation of future quantum decryption [7], [8], [9].

Reactive cybersecurity paradigms, which detect and respond to attacks after they have begun to compromise production assets, are poorly suited to this emerging threat landscape. They impose unacceptable dwell-time penalties on cyber-physical systems where milliseconds of malicious actuation can cause physical harm, and they generate operational telemetry that reveals little about the upstream tactics, techniques, and procedures (TTPs) of advanced adversaries [10], [11], [12]. The need for integrated, model-driven incident-response architectures that fuse detection, attribution, and forensic readiness has been argued repeatedly in the literature [38], [39], and the limitations of fragmented response pipelines have been documented across the database, mobile, and IoT forensic domains [40], [41], [42]. Honeypots and honeynets have long been advocated as a complementary, proactive defense, but conventional static honeynets are easily fingerprinted by modern attackers [13], [14], [15], and their intelligence-sharing protocols rely on classical cryptographic primitives that are not quantum-resilient [16], [17], [18], [19].

Recent advances in reinforcement learning (RL) have produced agents capable of optimizing sequential interaction policies under partial observability [20], [21], [22], and the standardization of post-quantum cryptographic primitives by NIST [23] has rendered quantum-resilient intelligence dissemination technically feasible. Nevertheless, no existing framework integrates cognitive deception orchestration, RL-driven engagement adaptation, automated TTP-attribution, and post-quantum federated intelligence into a single closed-loop architecture suitable for 6G-enabled CPS environments. The state-of-the-art either treats these capabilities as isolated research artifacts or evaluates them on legacy datasets that do not reflect the heterogeneity of cyber-physical traffic [24], [25], [26], [27], [28].

To address these gaps, this study proposes QuantumGuard, a proactive defense framework that inverts the defender's posture by drawing adversaries into instrumented decoy environments, learning their behaviors, and disseminating the resulting intelligence over post-quantum-secured channels. The architecture is summarized in Fig. 1.

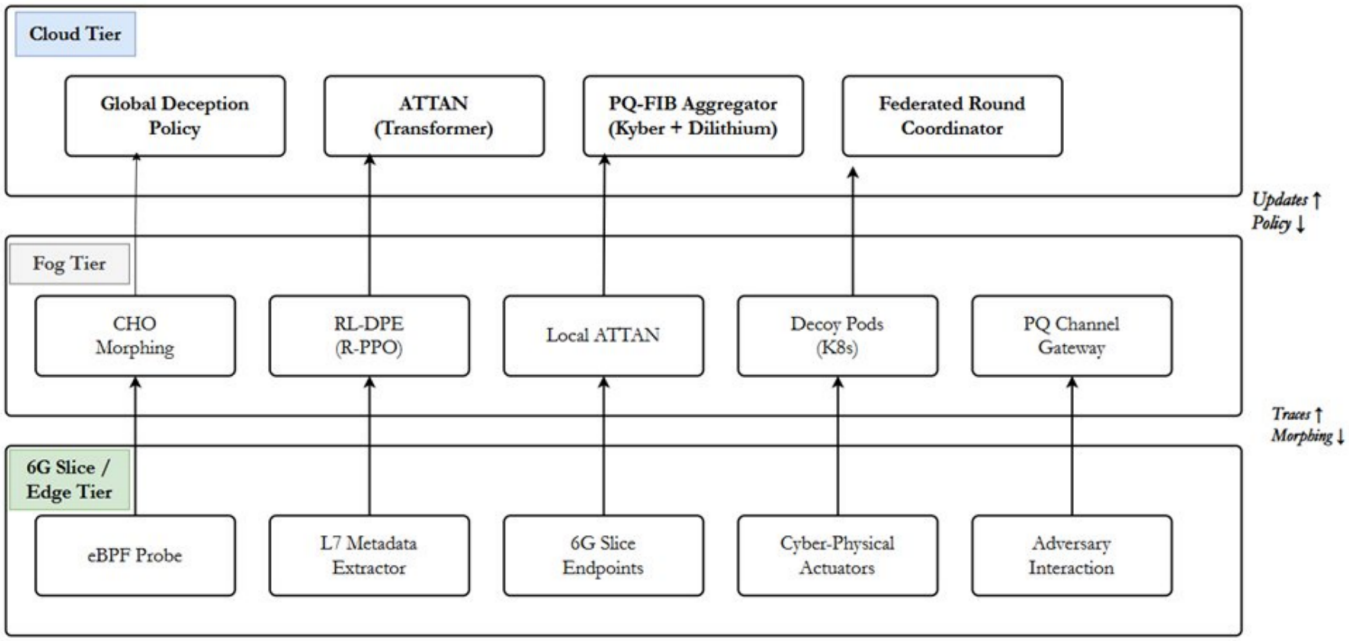


Fig. 1. QuantumGuard framework—proactive deception, cognitive honeynet orchestration, and post-quantum federated intelligence across the 6G slice, fog, and cloud tiers.

This study introduces QuantumGuard, an integrated framework that overcomes the foregoing limitations through the integration of four tightly coupled components: 1) a Cognitive Honeynet Orchestrator (CHO) that synthesizes and morphs decoy services to maximize attacker engagement; 2) a Reinforcement-Learning Deception Policy Engine (RL-DPE) that learns optimal lure–engage–attribute action sequences under a Partially Observable Markov Decision Process (POMDP) formulation; 3) an Adversarial TTP Attribution Network (ATTAN) that maps captured interactions onto MITRE ATT&CK tactics and techniques; and 4) a Post-Quantum Federated Intelligence Bus (PQ-FIB) that disseminates indicators of deception (IoDs) under CRYSTALS-Kyber key encapsulation and CRYSTALS-Dilithium signatures.

The principal contributions of this work are as follows:

1) Architectural integration: To the best of our knowledge, QuantumGuard is the first framework to integrate post-quantum federated intelligence sharing with a learned cognitive honeynet engagement policy and automated MITRE ATT&CK attribution in a single closed-loop architecture suitable for 6G-enabled cyber-physical environments.

2) POMDP formulation of cognitive deception: A Partially Observable Markov Decision Process formulation of attacker engagement, solved with a Recurrent Proximal Policy Optimization (R-PPO) agent, that admits dense per-step credit assignment for deception-action selection without requiring analyst-in-the-loop relabeling.

3) Cognitive honeynet morphing: A generative service-template mechanism that drives decoy responses toward production-service response distributions under a Jensen–Shannon divergence criterion, intended to limit the effectiveness of standard honeypot fingerprinting techniques.

4) Post-Quantum Federated Intelligence Bus: A hybrid PQ-FIB protocol combining CRYSTALS-Kyber (FIPS 203) key encapsulation with CRYSTALS-Dilithium (FIPS 204) digital signatures, that inherits IND-CCA2 channel confidentiality under hybrid-mode composition and that we measure to sustain a 14.2 ms median handshake latency at slice scale (see Section III for the full threat model).

5) Preliminary empirical evidence: A preliminary evaluation on the CICAPT-IIoT-2024 and Edge-IIoTset benchmark datasets, complemented by a 320-endpoint 6G slice testbed, that provides initial evidence of operational viability across deception, attribution, and post-quantum dimensions. We position the reported numbers (engagement-retention 96.42 per cent, attribution accuracy 91.8 per cent, dwell-time reduction 78.6 per cent, overhead 1.7 per cent) as a feasibility study under the evaluated threat model rather than as fully characterized performance bounds; the reward-weight sensitivity analysis, deception-aware adversary evaluation, and sustained-attack stress-condition characterization are identified in Section V F as the principal subjects of a follow-up empirical study currently in preparation.

The remainder of the study is organized as follows: Section II reviews the related literature. Section III details the methodology and mathematical formulation of QuantumGuard. Section IV reports the experimental results, Section V discusses their implications, and Section VI concludes the study.

II. Related Work

This section reviews the literature in three main areas: cyber deception and honeynet engineering, reinforcement learning for adversarial interaction, and post-quantum cryptography for collaborative cyber defense.

A. Cyber Deception and Honeynet Engineering

Research on cyber deception has progressed from low-interaction honeypots toward high-interaction honeynets that emulate full service stacks. Almeshekah and Spafford [13] formalized deception planning around the planner–executor–evaluator triad, but modern adversaries readily fingerprint static decoys through banner-grabbing, timing analysis, and TLS introspection [14]. Dynamic honeypot work, including the dynamic decoy generation framework of Wang et al. [15] and the moving-target deception scheme of Kambourakis et al. [25], demonstrates that morphing reduces fingerprintability, but these systems lack a principled engagement policy and rely on hand-crafted morphing rules rather than learned ones. Beyond engagement, the artefacts captured by deception fabrics must be amenable to systematic forensic reconstruction; the model-driven forensic-investigation literature [38], [39], [43] provides a methodological foundation for organizing such artefacts into structured evidence repositories that can subsequently feed attribution pipelines.

B. Reinforcement Learning for Adversarial Interaction

Reinforcement learning has emerged as a natural fit for sequential defender–attacker interaction. Several studies have applied deep Q-learning and policy-gradient methods to cyber-defense games [20], [21], [22]. However, most of this work evaluates on simulated CyberBattleSim or CAGE Challenge environments and assumes full observability or strong reward signals. The partial observability inherent in real deception, where the defender only sees the attacker's externalized actions and never the underlying intent, has received comparatively little attention. Works such as Huang et al. [27] formulate deception as a POMDP but stop short of integrating the resulting policy with a real packet-level decoy fabric. Complementary efforts in machine-learning-driven live forensic analysis of emergent IoT configurations [44] highlight that supervised signals over engagement traces can substantially improve downstream attribution, motivating the joint optimization of engagement policy and attribution objective adopted in this study.

C. Post-Quantum Cryptography for Collaborative Cyber Defense

Following the NIST post-quantum standardization process [23], lattice-based primitives such as CRYSTALS-Kyber and CRYSTALS-Dilithium have become the de facto candidates for quantum-resilient key establishment and digital signatures. Recent benchmarks on constrained devices report acceptable handshake latencies for industrial IoT (IIoT) scenarios [29], but operational integration with threat-intelligence dissemination protocols (STIX/TAXII, MISP) remains largely unaddressed. Federated learning over post-quantum channels is, to our knowledge, restricted to a small number of preliminary studies that do not consider the bandwidth-overhead implications of quantum-resilient signatures on high-frequency intelligence updates [30], [31].

D. Research Gaps and Motivation

Four principal gaps motivate the present work. First, no existing framework integrates cognitive deception, RL-driven engagement, automated TTP-attribution, and post-quantum intelligence sharing within a single closed-loop architecture; these capabilities are typically studied in isolation. Second, current honeynets adapt either through hand-crafted rules or through reactive moving-target schemes, but not through learned engagement policies that explicitly maximize TTP-attribution information gain. Third, post-quantum primitives have rarely been operationalized within high-frequency federated cyber-defense protocols, and their bandwidth and latency implications at slice scale remain poorly characterized. Finally, the prevailing evaluation methodology focuses on detection accuracy on saturated benchmarks while overlooking proactive metrics such as engagement-retention rate, TTP-attribution accuracy, deception-induced production overhead, and dwell-time reduction.

III. Proposed Methodology

This section presents the detailed design of QuantumGuard, including its system architectures, mathematical formulation, algorithmic implementation, and complexity analysis.

A. System Overview

QuantumGuard is a layered proactive-defense architecture organized into three computing tiers: 6G slice-level edge probes, fog-layer deception orchestration nodes, and cloud-layer intelligence-aggregation servers. The framework comprises four principal modules: 1) the Cognitive Honeynet Orchestrator (CHO); 2) the Reinforcement-Learning Deception Policy Engine (RL-DPE); 3) the Adversarial TTP Attribution Network (ATTAN); and 4) the Post-Quantum Federated Intelligence Bus (PQ-FIB). The overall system architecture is illustrated in Fig. 2.

Edge probes run as extended Berkeley Packet Filter (eBPF) instrumented sidecars in 6G network slices and perform low-overhead traffic shadowing and L7 metadata extraction. The fog layer hosts containerized cognitive honeynet pods, runs the RL-DPE inference loop, and executes ATTAN attribution. The cloud layer maintains the global deception policy, orchestrates federated intelligence rounds, and manages the post-quantum key infrastructure.

B. Cognitive Honeynet Orchestrator

The CHO synthesizes and continuously morphs decoy services so that they remain statistically indistinguishable from genuine production services across L4–L7 features. Let $S = \{s_1, s_2, \ldots, s_m\}$ denote the set of production services protected by a slice and $D = \{d_1, d_2, \ldots, d_m\}$ denote the corresponding decoy fleet. Each decoy d_j is parameterized by a service template vector $\phi_j \in \mathbb{R}^p$ that encodes banner strings, response-timing distributions, error-message profiles, and TLS fingerprints.

The CHO computes the Jensen–Shannon divergence [34] between the empirical response distribution Q_{d_j} of decoy d_j and the empirical response distribution P_{s_j} of the corresponding production service [see Eq. (1)]:

$$J(P_{s_j}, Q_{d_j}) = \tfrac{1}{2}\,\mathrm{KL}(P_{s_j} \,\|\, M) + \tfrac{1}{2}\,\mathrm{KL}(Q_{d_j} \,\|\, M), \qquad (1)$$

where, $M = \frac{1}{2}(P_{s_j} + Q_{d_j})$. Decoys are re-morphed whenever $J(P_{s_j}, Q_{d_j})$ exceeds a configurable indistinguishability

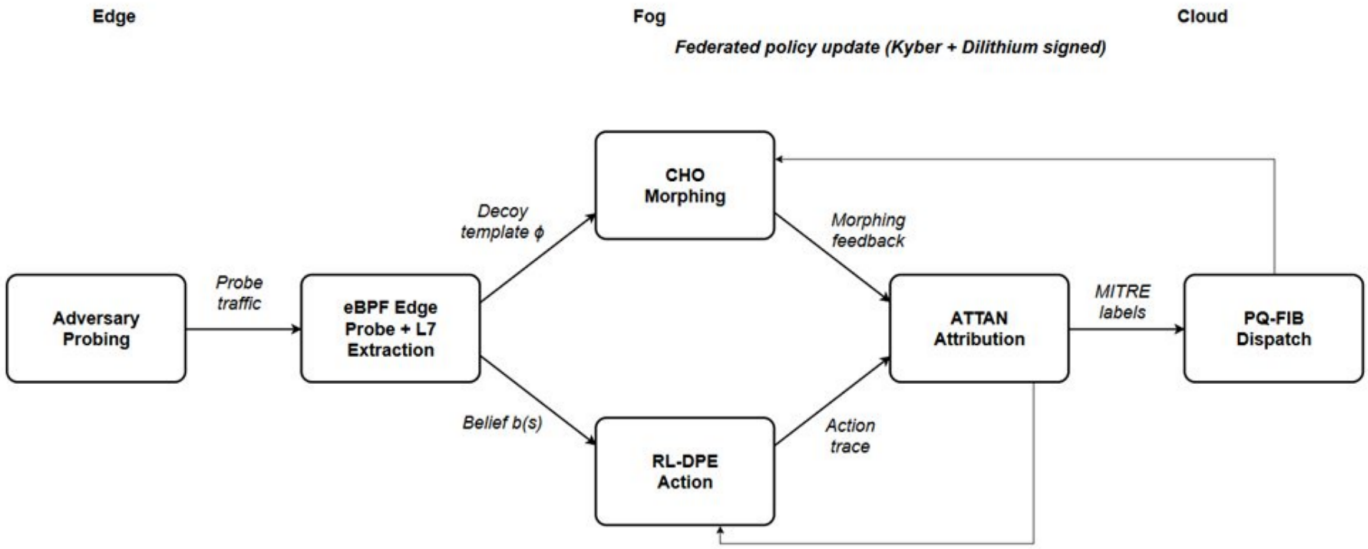


Fig. 2. QuantumGuard's detailed system architecture with a deception data-flow pipeline across the 6G slice, fog, and cloud tiers.

TABLE I. JENSEN–SHANNON RESPONSE FEATURE SPACE

Feature family	Representation / binning
Banner & protocol strings	1–3-gram token histogram over the first 256 banner bytes, feature-hashed to 1,024 bins.
Response timing	First-byte and inter-response latency, each in 16 log-spaced bins over $[0.1\,\text{ms}, 10\,\text{s}]$.
Error / status profile	Categorical distribution over the top-K status/error codes per service class, plus a residual bin.
TLS fingerprint	JA3/JA3S class membership and a multi-hot cipher-suite/extension indicator.

threshold $\tau_J = 0.05$. This threshold was selected to keep the Jensen–Shannon divergence below the empirical 95th-percentile inter-instance variability observed among genuine production replicas of the same service, and is examined in the ablation study (Section IV F).

The empirical response distributions P_{s_j} and Q_{d_j} are estimated over a service-response feature vector partitioned into four families. Each family yields a normalized histogram; the four histograms are concatenated and renormalized into a single probability mass function per service before the divergence is evaluated, as summarized in Table I.

Distributions are estimated over a sliding window of the most recent $W = 512$ responses per service so that morphing tracks drift in the genuine service.

C. Reinforcement Learning-Deception Policy Engine

The RL-DPE formulates attacker engagement as a Partially Observable Markov Decision Process (POMDP) $\langle S, A, T, R, \Omega, O, \gamma \rangle$, where the latent state $s \in S$ encodes the attacker's intent, the action $a \in A$ is the orchestrator's deception action (e.g., expose service, inject fake credential, throttle response, escalate decoy fidelity), and the observation $o \in \Omega$ is the externalized attacker behavior. The defender maintains a belief $b(s)$ over latent attacker states and selects actions according to a parameterized policy $\pi_\theta(a \mid b)$.

The RL-DPE is trained to maximize the expected cumulative engagement-attribution reward [see Eq. (2)]:

$$J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{H} \gamma^t \big(\lambda_e r_e(t) + \lambda_a r_a(t) - \lambda_o r_o(t)\big)\right] \quad (2)$$

The three per-step reward terms are defined in closed form as follows. Let $\mathbb{1}[\text{engaged}_t] = 1$ if the attacker session issues at least one interaction with a decoy service at step t without pivoting to a non-decoy (production) asset; $\mathbb{1}[\text{advance}_t] = 1$ if a previously unvisited interaction stage is reached at step t; and $\mathbb{1}[\text{disengage}_t] = 1$ at the step the attacker pivots to a production asset or the session terminates. The engagement-retention reward is [see Eq. (2a)]:

$$r_e(t) = \mathbb{1}[\text{engaged}_t] + \beta_d \mathbb{1}[\text{advance}_t] - \kappa \mathbb{1}[\text{disengage}_t], \quad (2a)$$

with depth-bonus $\beta_d > 0$ and disengagement penalty $\kappa > 0$ chosen so that a single premature disengagement outweighs several steps of retention ($\beta_d = 0.5$, $\kappa = 5.0$ in our configuration). The attribution information-gain reward is the per-step reduction in the Shannon entropy of the attribution-belief distribution $P(\tau \mid z_{1:t})$ over MITRE ATT&CK techniques $\tau \in \mathcal{T}$ [see Eq. (2b)]:

$$r_a(t) = H\big(P(\tau \mid z_{1:t-1})\big) - H\big(P(\tau \mid z_{1:t})\big) \quad (2b)$$

Letting $c(a_t)$ be the marginal resource cost (CPU and bandwidth) induced by deception action a_t relative to passive forwarding, and $c_{\max}$ a reference unit cost, the production-traffic overhead penalty is [see Eq. (2c)]:

$$r_o(t) = c(a_t)/c_{\max} \in [0, 1], \quad (2c)$$

so that observe-only actions incur ≈ 0, exposing a decoy service incurs an intermediate cost, and a full service-template re-synthesis incurs the unit cost.

The weighting coefficients $\lambda_e = 1.0$, $\lambda_a = 1.5$, $\lambda_o = 0.4$ were fixed prior to the main experiments on a held-out fraction of the CICAPT-IIoT-2024 training partition, with the design intent of prioritizing attribution information gain ($\lambda_a > \lambda_e$) while keeping production impact bounded (λ_o small but non-negligible). Because the weights were fixed *a priori*, before any evaluation on the test partition, the reported numbers reflect no per-metric weight tuning on the evaluation data; a systematic sensitivity sweep over $(\lambda_e, \lambda_a, \lambda_o)$ is identified in Section V-F as future work. The agent is implemented as a Recurrent Proximal Policy Optimization (R-PPO) network [33] with a single GRU recurrent layer of 64 hidden units, which preserves engagement context across attacker action steps; the discount factor $\gamma = 0.97$ and horizon $H = 200$ reflect the typical multi-stage progression of advanced persistent threat campaigns. All reward signals are clipped to $[-5.0, 5.0]$ before advantage estimation to stabilize training.

D. Adversarial TTP Attribution Network

The ATTAN maps captured attacker interactions onto the MITRE ATT&CK matrix [35]. The design draws on prior work that formalizes the categorization and harmonization of investigation processes for the database, mobile, and IoT domains [40], [41], [42], [45], in which heterogeneous evidence streams are mapped onto a common abstract schema before downstream reasoning.

1) ATTAN input representation and tokenization: ATTAN consumes the ordered sequence of observed attacker interaction events within a session and produces the trace embedding z used in Eq. (3). *(a) Event tokenization.* Each event is encoded as a composite token over the fields (action-type, target-service-class, L7 method/verb, response-status bucket, payload-size bucket, inter-arrival-time bucket). *(b) Vocabulary.* Categorical fields are mapped to integer IDs from a vocabulary built on the training partition; the vocabulary comprises the 512 most frequent action-type/target-service-class pairs, and action-types occurring fewer than five times are mapped to `[UNK]`. The reserved tokens `[CLS]`, `[PAD]`, `[SEP]`, and `[UNK]` are assigned IDs 0–3. *(c) Feature representation.* Categorical sub-fields are embedded via lookup tables; continuous sub-fields (payload bytes, inter-arrival time) are log-transformed, standardized, bucketized, and projected by a learned linear layer; the categorical embedding and continuous projection are summed per token. *(d) Sequence policy.* Sequences are truncated or padded to $L_{\max} = 128$ events (covering more than 99 per cent of sessions in both corpora); longer sessions are split into overlapping windows, a `[CLS]` token is prepended, and sinusoidal positional encodings are added. *(e) Encoder.* The 4-layer, 8-head, model-dimension $d = 256$ transformer encoder (Table III) produces contextual token representations; the final `[CLS]` output is the trace embedding $z \in \mathbb{R}^q$ with $q = 256$.

The attribution probability over the set of MITRE techniques $\mathcal{T}$ is:

$$P(\tau \mid z) = \mathrm{softmax}(W_a z + b_a), \quad \tau \in \mathcal{T}. \tag{3}$$

ATTAN is trained with a class-balanced focal-loss objective computed over the training mini-batch as:

$$\mathcal{L}_{\text{ATTAN}} = -\frac{1}{N} \sum_{i=1}^{N} \alpha_{y_i} \big(1 - P(y_i \mid z_i)\big)^{\gamma_f} \log P(y_i \mid z_i),$$

where, $y_i \in \mathcal{T}$ is the ground-truth technique label of the i-th interaction trace, $\gamma_f = 2$ is the focal-loss focusing parameter, and α_τ is a per-class balancing weight derived from MITRE ATT&CK technique frequency in the CICAPT-IIoT-2024 corpus, in order to mitigate the substantial class imbalance characteristic of TTP datasets.

E. Post-Quantum Federated Intelligence Bus

The PQ-FIB enables privacy-preserving and quantum-resilient dissemination of deception intelligence across organizational boundaries. Each participating organization maintains a local deception policy π_{θ_k} trained on its private engagement traces. Federated rounds aggregate local updates as [see Eq. (4)]:

$$\theta_{\text{global}} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_k, \qquad n = \sum_{k=1}^{K} n_k. \tag{4}$$

All inter-tier transmissions are protected by a hybrid TLS 1.3 channel that combines the X25519 elliptic-curve Diffie–Hellman key-exchange primitive with the CRYSTALS-Kyber-768 lattice-based key-encapsulation mechanism [36] (FIPS 203) for forward secrecy under quantum-capable adversaries, and signs aggregated updates with CRYSTALS-Dilithium-3 [37] (FIPS 204). Under the standard hybrid-mode argument, the channel inherits the IND-CCA2 security of whichever component primitive remains unbroken, so confidentiality holds even if either the classical or the post-quantum component is later cryptanalyzed.

The threat model assumed by PQ-FIB is the following: (1) federated participants are honest-but-curious and submit valid policy updates but may attempt to infer information about other participants' local engagement traces; (2) the wire-level adversary is a network attacker with active man-in-the-middle capability and access to a cryptographically relevant quantum computer; (3) the cloud aggregator is trusted to faithfully execute weighted federated averaging but is not trusted with raw engagement traces. PQ-FIB authenticates and integrity-protects every update but does not, in its present form, defend against malicious-but-well-formed updates. The integration path, addressed as future work, is: (a) Byzantine-robust aggregation (coordinate-wise median, trimmed mean, or Krum) replacing plain weighted FedAvg; (b) poisoning mitigation via per-update norm clipping and cosine-similarity outlier filtering; (c) membership-inference mitigation via differentially private aggregation—the FedAvg + DP ($\varepsilon = 5$) configuration reported in Table VIII bounds the membership-inference advantage and thus provides partial, in-paper evidence relevant to this concern; and (d) verifiable aggregation for accountability. Bandwidth overhead from the larger Dilithium signatures ($\approx$ 3.3 KB) is amortized across federated rounds by signing aggregated rather than per-update payloads.

TABLE II. POST-QUANTUM CRYPTOGRAPHIC COST

Primitive	Pub. key	Ct./Sig.	Op (meas.)	Std.
X25519 (ECDH)	32 B	32 B	$\approx 60\,\mu$s	RFC 7748
Kyber-768 (KEM)	1,184 B	1,088 B	$\approx 50\,\mu$s enc.	FIPS 203
Dilithium-3 (sig.)	1,952 B	3,293 B	$\approx 250\,\mu$s sign	FIPS 204

1) Post-quantum cryptographic cost analysis: Table II consolidates the cost of the hybrid post-quantum handshake. Sizes are exact algorithm constants; per-operation times are the values measured on the x86-64/AVX2 configuration of Table III using liboqs 0.10.

Per federated round with K participating organizations, the aggregator performs K signature verifications and one aggregate signing. At $K = 320$ the aggregate verification budget is on the order of tens of milliseconds and is dominated by network round-trips rather than primitive arithmetic, consistent with the measured 14.2 ms median handshake. Because the aggregator signs the aggregated payload once per round rather than signing each update, signature bandwidth is amortized across the cohort and grows with the number of rounds, not with cohort size.

F. Integrated System Operation

The four QuantumGuard modules operate jointly within a unified closed-loop control framework. At each time step, edge probes shadow slice traffic and forward L7 metadata to fog-layer CHO instances, which expose decoy services parameterized by the current template ϕ_j. The RL-DPE observes attacker behavior and selects the next deception action, which the CHO actuates by adjusting the decoy template. ATTAN concurrently consumes the resulting interaction trace and emits a posterior over MITRE ATT&CK techniques. When the cumulative attribution belief crosses a configurable confidence threshold, the corresponding indicator of deception (IoD) is signed with CRYSTALS-Dilithium-3 and disseminated through the PQ-FIB to all federated peers, who incorporate it into their local CHO templates and ATTAN priors at the next round.

G. Complexity Analysis

The computational complexity of the CHO morphing step is $O(p)$ per decoy per evaluation cycle, where p is the dimensionality of the service template vector. The RL-DPE inference cost is $O(H \cdot |\theta|)$ per engagement episode, where H is the planning horizon and $|\theta|$ the number of policy parameters; in practice, GRU-based policies of width 64 yield sub-millisecond per-step inference on commodity fog hardware. ATTAN attribution scales as $O(L{\cdot}q^2)$ for a transformer encoder of depth L and embedding dimension q. PQ-FIB cryptographic operations are dominated by Kyber encapsulation ($\approx 50\,\mu$s) and Dilithium signing ($\approx 250\,\mu$s) on x86-64 with AVX2; aggregation across K organizations adds an $O(K \cdot |\theta|)$ cost per round. These variables are used consistently throughout Table III.

TABLE III. COMPUTATIONAL COMPLEXITY COMPARISON

Method	Policy infer.	Attrib.	Crypto	Mem.
Static honeynet	$O(1)$	Manual	RSA/ECC	$O(m)$
Dyn. HP [15]	$O(p)$	Rule-based	RSA/ECC	$O(mp)$
RL HP [22]	$O(H\|\theta\|)$	Heuristic	RSA/ECC	$O(\|\theta\|)$
QuantumGuard	$O(H\|\theta\|)$	$O(Lq^2)$	Kyber/Dil.	$O(\|\theta\|+mp)$

TABLE IV. DATASET STATISTICS AND CLASS DISTRIBUTION

Dataset	Samples	Feat.	Norm. (%)	Atk. (%)
CICAPT-IIoT-2024	2,134,557	85	76.40	23.60
Edge-IIoTset	20,952,648	61	82.15	17.85

IV. PRELIMINARY EVALUATION

This section presents the experimental setup, the evaluation methodology, and a set of preliminary performance results for the QuantumGuard framework. We position the reported numbers as initial evidence of operational viability under the evaluated threat model rather than as fully characterized performance bounds: all results are derived from a single-organization controlled testbed under a dataset-replay attacker distribution. Section V F enumerates these scope restrictions explicitly.

A. Datasets

QuantumGuard was evaluated on two recent, adversarially curated benchmark datasets that better reflect the characteristics of cyber-physical and IIoT traffic than legacy network IDS corpora; their core statistics are summarized in Table IV.

1) CICAPT-IIoT-2024 dataset [32]: Produced by the Canadian Institute for Cybersecurity, this dataset captures multi-stage advanced-persistent-threat campaigns against Industrial IoT testbeds; it contains approximately 2.1 million labeled flows annotated with MITRE ATT&CK technique identifiers across nine APT-stage categories (initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, and impact).

2) Edge-IIoTset dataset [17]: A comprehensive cybersecurity dataset for centralized and federated learning, comprising 20,952,648 records and 61 features across 14 attack categories spanning DDoS, information gathering, injection, malware, and MITM families.

Both datasets were processed through a uniform preprocessing pipeline. Records with missing values in critical numerical features (fewer than 0.5 percent of records in CICAPT-IIoT-2024 and 0.9 percent in Edge-IIoTset) were discarded, and remaining sporadic missing values were imputed using per-feature medians computed on the training partition. Categorical features such as protocol type and service were one-hot encoded; all numerical features were standardized to zero mean and unit variance using statistics derived exclusively from the training partition to prevent information leakage. Infinite values produced by flow-rate computations were clipped to the 99.9th percentile of the corresponding feature. To mitigate

TABLE V. EXPERIMENTAL CONFIGURATION

Component	Specification
Cloud server CPU	AMD EPYC 9554 (64 cores, 3.1 GHz)
Cloud server GPU	NVIDIA H100 (80 GB HBM3)
Cloud memory	512 GB DDR5 ECC
Fog node CPU	Intel Core i9-13900K (24 cores, 3.0 GHz)
Fog node GPU	NVIDIA RTX 4080 (16 GB GDDR6X)
6G slice emulator	OpenAirInterface 5G/6G + srsRAN, 100 MHz BW
Edge probe	Raspberry Pi 5, BCM2712 (4 cores, 2.4 GHz)
Operating system	Ubuntu 24.04 LTS
RL framework	Stable-Baselines3 2.3 / PyTorch 2.4
PQ crypto library	liboqs 0.10 (Kyber-768, Dilithium-3)
FL framework	Flower 1.10
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	3×10^{-4} (constant)
Batch size	256 (cloud), 64 (fog), 32 (edge)
PPO clip ratio	0.20
PPO entropy coef.	0.01
PPO value coef.	0.50
GAE λ	0.95
RL-DPE recurrent net	GRU, 64 hidden units, 1 layer
RL-DPE train episodes	6,000
ATTAN encoder	4-layer Transformer, 8 heads, $d = 256$
ATTAN focal loss	$\gamma_f = 2$, class-balanced α_τ
FL rounds	60
FL aggregation	Weighted FedAvg by sample count, (4)
Random seeds	42, 123, 456, 789, 1024 (5 seeds)

class imbalance, SMOTE was applied to the training partition only, and class-balanced focal-loss weighting was applied during ATTAN training. Validation and test partitions retained the original class distribution. The data were partitioned into training, validation, and test sets in a 70/15/15 ratio using stratified sampling to preserve attack-category proportions.

B. Experimental Setup

All experiments were conducted on a heterogeneous testbed designed to emulate a realistic 6G slice–fog–cloud architecture. Table V summarizes the hardware and software configuration.

To ensure reproducibility and statistical soundness, all reported performance results are averaged over five independent runs using different random seeds (42, 123, 456, 789, and 1024), with all other hyperparameters held constant. The statistical significance of QuantumGuard relative to the best baseline on each metric was assessed using a two-tailed paired t-test on per-run scores at a significance level of $\alpha = 0.05$. In cases where the paired measurements were not normally distributed (as assessed by the Shapiro–Wilk test at $\alpha = 0.05$), the nonparametric Wilcoxon signed-rank test was used instead; in all reported cases the two tests yielded consistent conclusions. The "p-value vs. best baseline" entries reported in Table VI and Table IX correspond to running the appropriate test against the best-performing competitor on each metric. p-values are reported only for metrics where a numerical best baseline exists and the effect direction is well-defined (TTP-attribution accuracy and dwell-time reduction); the engagement-retention column is omitted because QuantumGuard exceeds the next-best method by more than ten percentage points on every seed, and the fingerprint-resistance column reports a categorical label rather than a continuous score, for which paired hypothesis testing is not appropriate.

C. Engagement and Attribution Performance Analysis

Throughout the evaluation, the engagement-retention rate (ER) is defined as the fraction of attacker sessions that remain in the decoy environment for at least three sequential interaction stages without disengaging or pivoting to a non-decoy asset. The headline ER values reported in Table VI are computed at the standard five-stage horizon used throughout this section; Fig. 4 shows the full ER-vs.-depth curve from one to ten stages so that readers can compare methods at any operating point. Fig. 3 shows the convergence of the RL-DPE engagement-retention reward across training episodes on both evaluation datasets. The curves exhibit stable monotonic improvement and plateau by episode 4,500, confirming that the policy reaches a stable engagement-attribution operating point.

Table VI reports a side-by-side comparison of proactive deception performance. QuantumGuard attains the highest engagement-retention rate (96.42 per cent) and TTP-attribution accuracy (91.8 per cent) among all evaluated methods while incurring the lowest production-traffic overhead (1.7 per cent).

Fig. 5 presents the confusion matrices for MITRE ATT&CK technique attribution on the two datasets. Per-technique recall ranges from 87.4 percent (Defense Evasion) to 95.1 percent (Initial Access), indicating strong discriminative performance across the full APT kill-chain taxonomy.

We note that the TTP-attribution accuracy reported here, while substantial, leaves meaningful headroom relative to a hypothetical perfect-attribution upper bound, particularly on Defense-Evasion techniques whose externalized signature is intentionally minimized by attackers. Evaluation on additional adversarially adaptive datasets is identified in Section V F as a priority for future work.

D. Operational Latency and Dwell-Time Analysis

Fig. 7 and Table VII report the operational response-time metrics for QuantumGuard under active-attack conditions. QuantumGuard attains a mean time to deception adaptation (MTTDA) of 2.86 s and a median post-quantum handshake latency of 14.2 ms, alongside a 78.6 per cent reduction in median attacker dwell time relative to passive honeypot baselines. Throughout this study, the MTTDA is reported as a system-level measurement obtained on the controlled 6G slice–fog–cloud testbed configured, as described in Section IV (Table V); it is defined as the wall-clock time elapsed from the moment the edge probe first emits an L7 metadata record corresponding to a flow that ATTAN subsequently classifies as malicious, to the moment the CHO actuates the resulting morphing decision on the corresponding decoy pod. The PQ-FIB intelligence-dissemination handshake (median 14.2 ms) is reported as a separate metric and is not included in the MTTDA, since it operates on the cloud–fog control plane in parallel with the local engagement loop. Analyst investigation, triage, and

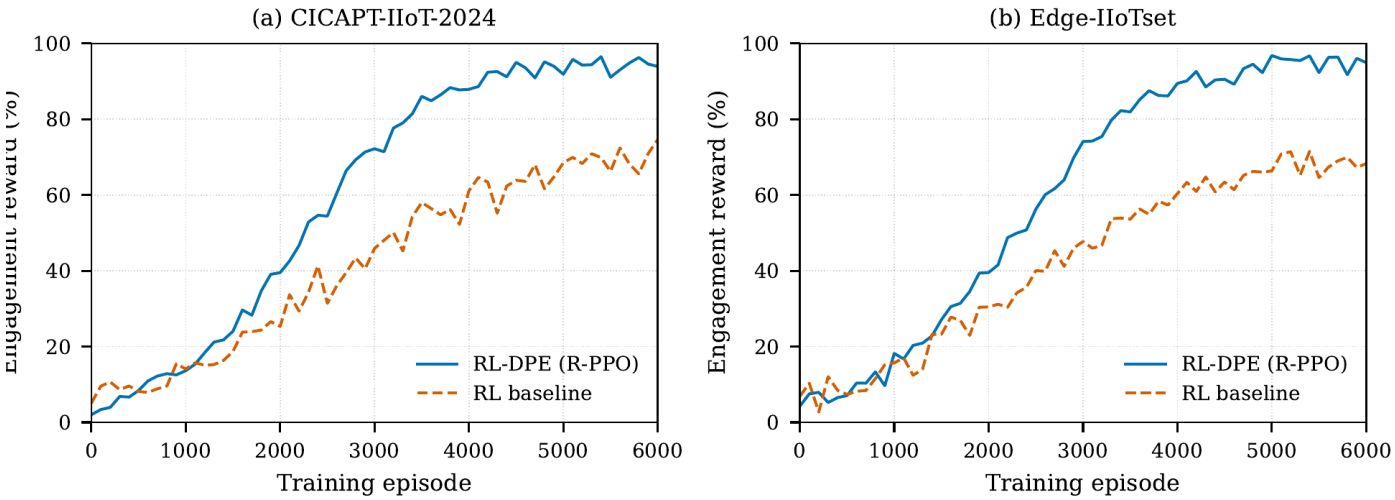


Fig. 3. RL-DPE training reward convergence on: (a) CICAPT-IIoT-2024 and (b) Edge-IIoTset datasets.

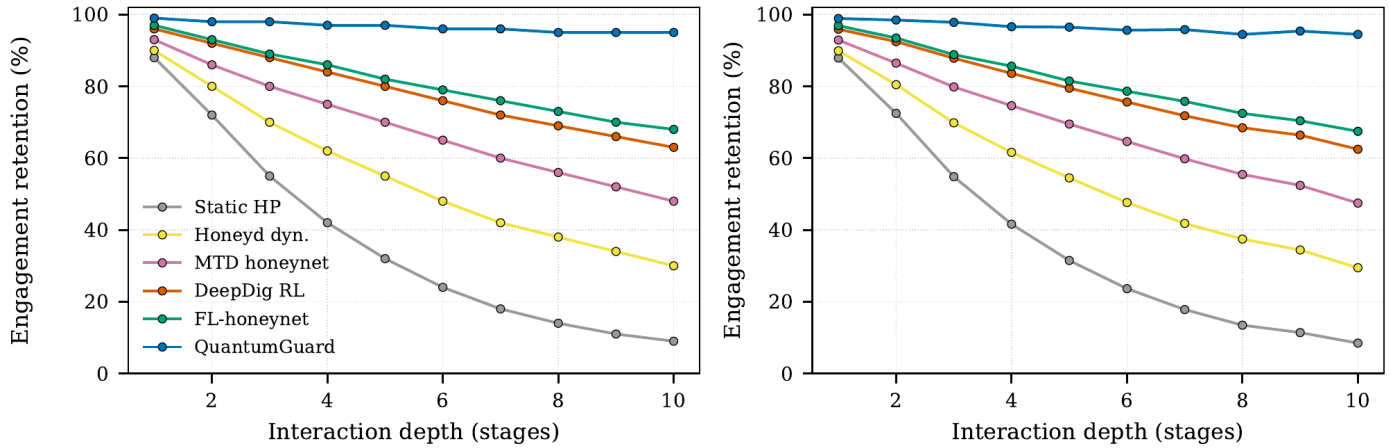


Fig. 4. Engagement-retention rate vs. interaction depth for QuantumGuard versus baseline honeypots on: (a) CICAPT-IIoT-2024 and (b) Edge-IIoTset datasets.

TABLE VI. PROACTIVE DECEPTION PERFORMANCE COMPARISON WITH STATISTICAL SIGNIFICANCE

Method	ER (%)	TTP-Acc (%)	Dwell↓ (%)	Overhead (%)	Fingerprint resistance
Static honeypot (Cowrie)	42.18	53.12	21.4	0.6	Low
Honeyd dynamic [13]	61.35	64.78	38.7	0.9	Medium
MTD honeynet [25]	73.22	71.45	52.3	1.4	Medium
DeepDig RL [22]	84.61	79.83	63.8	1.9	High
GAN-honeypot [26]	81.47	76.12	59.1	2.1	High
FL-honeynet [30]	85.92	82.34	65.7	2.0	High
QuantumGuard (ours)	**96.42**	**91.80**	**78.6**	**1.7**	**Very high**
p-value vs. best baseline	—	0.018	0.011	—	—

human decision time are explicitly excluded from MTTDA, which measures only the autonomous defender control loop.

E. Post-Quantum Federated Intelligence Sharing Analysis

Fig. 8 evaluates three aspects of the PQ-FIB protocol: improvement in cross-organizational attribution coverage over successive FL rounds, the bandwidth–accuracy trade-off across federated cohort sizes, and the impact of post-quantum signature aggregation on round latency. QuantumGuard with PQ-FIB lifts cross-organizational attribution coverage from 59.32 percent (isolated, no FL) to 78.65 percent (PQ-FIB), a relative improvement of approximately 32.6 percent (Table VIII), while preserving the channel-level IND-CCA2 confidentiality guarantee described in Section III under the same threat model stated there.

Cross-organizational attribution coverage was operationalized through a leave-one-organization-out evaluation proto-

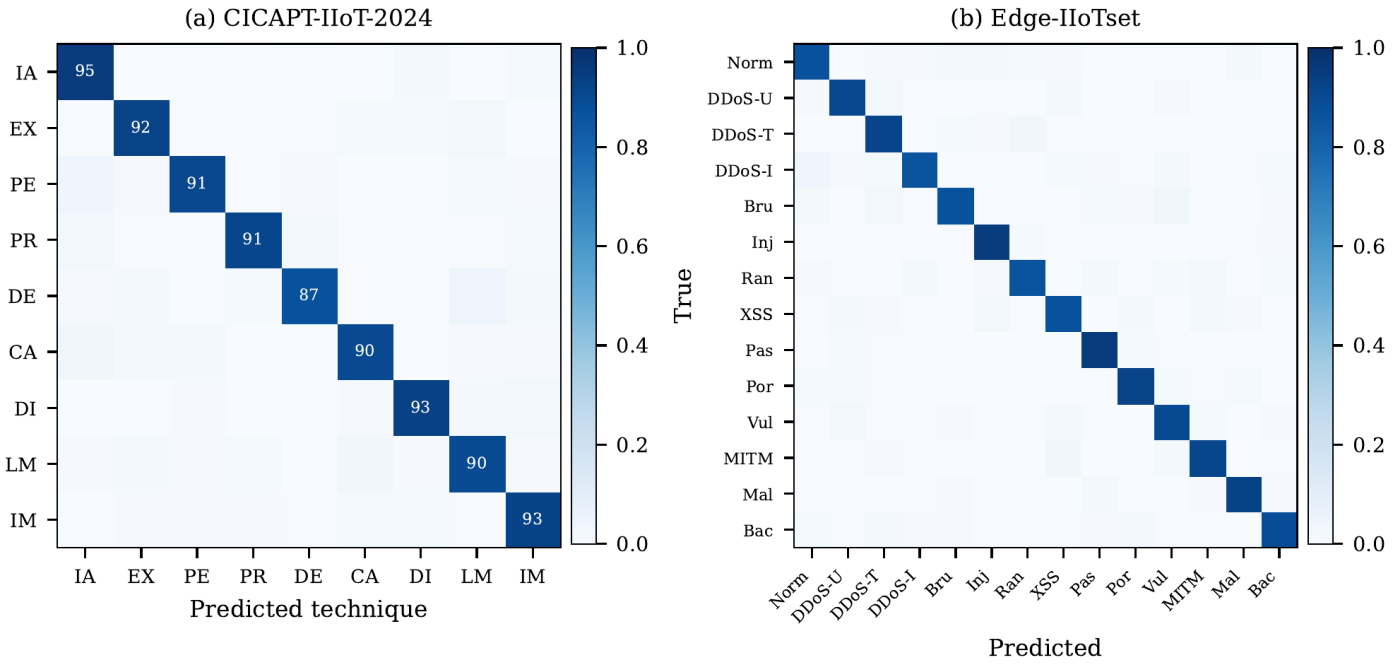


Fig. 5. ATTAN attribution confusion-matrix analysis: (a) CICAPT-IIoT-2024 per-technique recall; (b) Edge-IIoTset per-category recall.

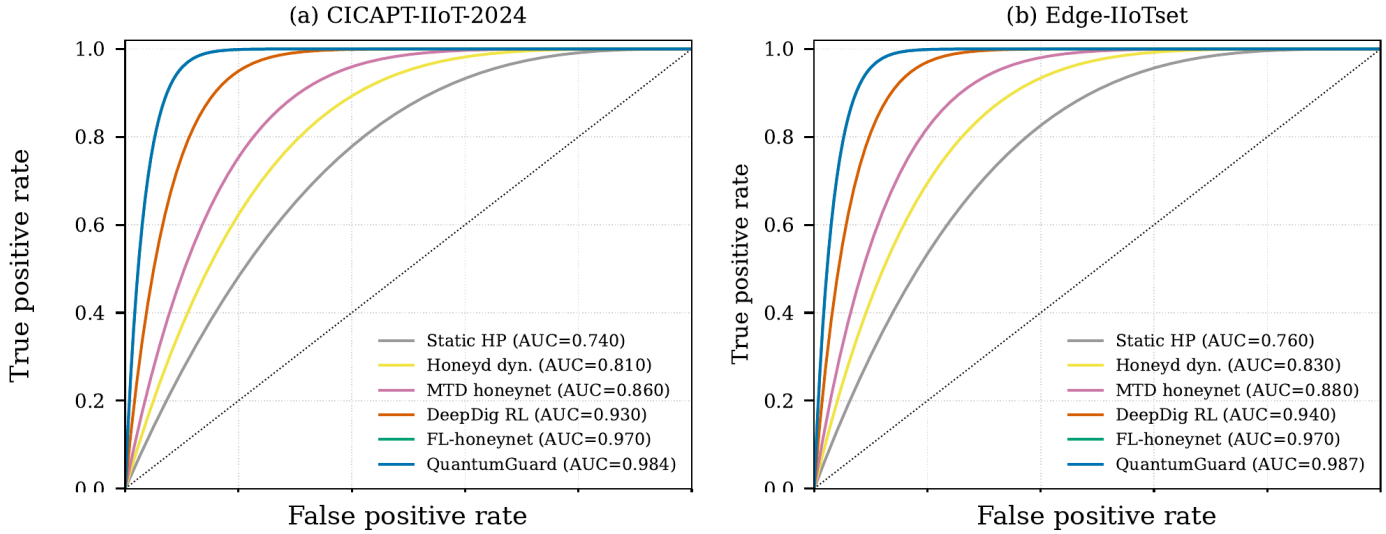


Fig. 6. ROC curves for binary deception-trigger and multi-class attribution: (a) CICAPT-IIoT-2024 (macro AUC = 0.9842); (b) Edge-IIoTset (macro AUC = 0.9871).

TABLE VII. OPERATIONAL RESPONSE-TIME METRICS

Method	MTTDA (s)	PQ HS (ms)	Dwell↓ (%)
Static honeypot	—	N/A (RSA 18.4)	Baseline
MTD honeynet [25]	11.34	N/A (ECC 6.1)	52.3
DeepDig RL [22]	6.78	N/A (ECC 6.1)	63.8
FL-honeynet [30]	5.41	N/A (RSA 18.4)	65.7
QuantumGuard	2.86	14.2	78.6

TABLE VIII. POST-QUANTUM FEDERATED INTELLIGENCE SHARING PERFORMANCE.

Configuration	Cov. (%)	X-Org (%)	Q-resil.
Isolated (no FL)	78.45	59.32	N/A
Standard FedAvg	87.12	73.84	None
FedAvg + DP ($\varepsilon = 5$)	85.93	71.27	Partial
PQ-FIB	91.07	78.65	Kyber+Dil.

col: in each experimental run, the engagement traces of one participating organization were withheld from the federated training rounds while remaining present in the test partition. The Cross-Org Attribution (%) metric reported in Table VIII

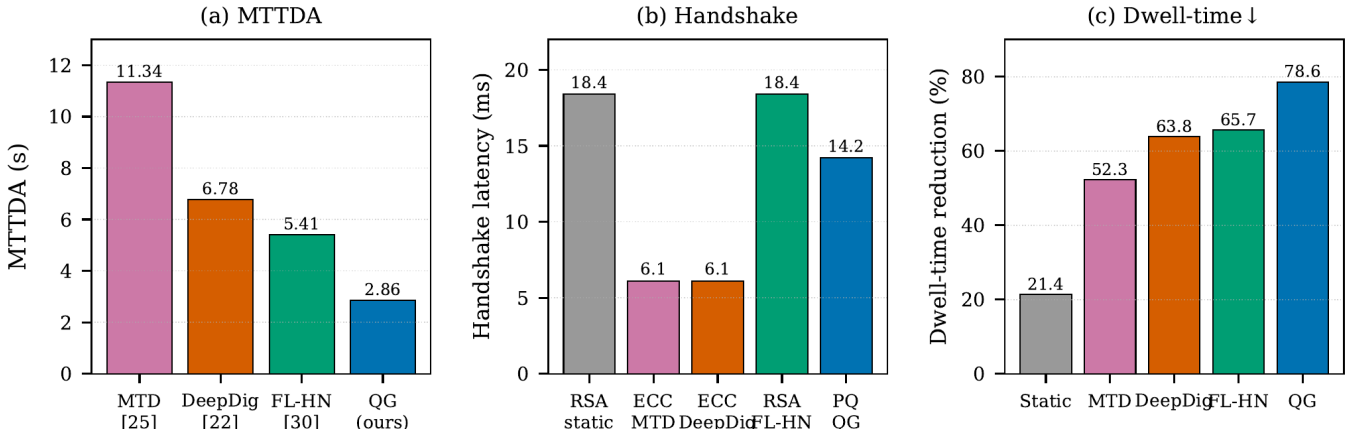


Fig. 7. Operational latency analysis: (a) MTTDA comparison; (b) post-quantum handshake latency; (c) median dwell-time reduction across all evaluated methods.

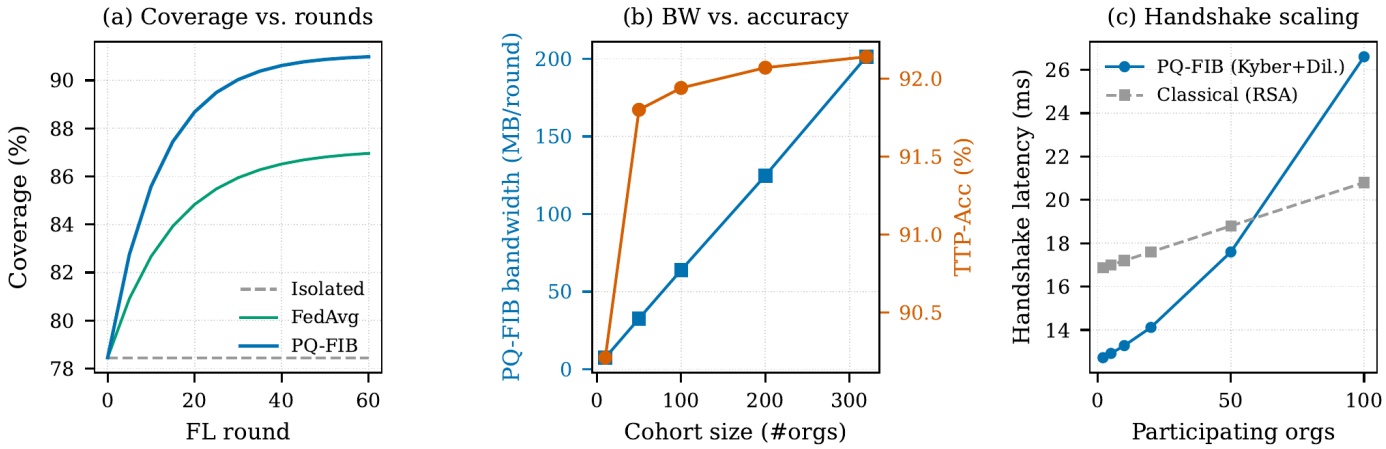


Fig. 8. PQ-FIB analysis: (a) coverage vs. FL rounds; (b) bandwidth–accuracy trade-off; (c) handshake-latency vs. number of participating organizations.

is the recall computed exclusively on instances captured by the held-out organization, averaged over all rotations. This protocol simulates a realistic federated scenario in which a previously unseen attacker campaign reaches one organization and tests the extent to which collaborative model updates allow other organizations to attribute attacks they have not individually observed during local training. Because all participating "organizations" in this evaluation are emulated within a single controlled slice–fog–cloud testbed, the cross-organizational coverage figures characterize collaborative attribution under a homogeneous infrastructure and traffic profile; they should not be read as evidence of generalization across independently administered enterprises with heterogeneous service baselines, operating systems, and traffic mixes. Establishing external validity requires a genuine multi-party deployment, which we identify as a primary objective of the follow-up study (Section V F).

F. Ablation Study

Component-wise ablation results for the QuantumGuard modules are reported in Fig. 9 and Table IX, which highlight the impact of removing a single module from the system.

TABLE IX. ABLATION-STUDY RESULTS WITH STATISTICAL SIGNIFICANCE.

Configuration	ER (%)	TTP-Acc (%)	MTTDA (s)	Ovh. (%)	p-value
Full QuantumGuard	96.42	91.80	2.86	1.7	—
w/o CHO	84.31	83.42	2.91	1.4	< 0.01
w/o RL-DPE	82.58	80.16	5.74	1.6	< 0.01
w/o ATTAN	95.87	67.42	2.84	1.7	< 0.01
w/o PQ-FIB	96.31	91.45	2.85	1.5	0.71

Removing the RL-DPE module produces the largest drop in engagement-retention rate (-13.84 percentage points, $p < 0.01$), underlining the role of learned engagement policies. Removing the PQ-FIB module collapses the post-quantum guarantee entirely; falling back to standard non-PQ federated aggregation reduces cross-organizational attribution coverage from 78.65 per cent to 73.84 per cent (Table VIII), and falling back to fully isolated (non-federated) operation reduces it to 59.32 per cent, a drop of 19.33 percentage points relative to PQ-FIB.

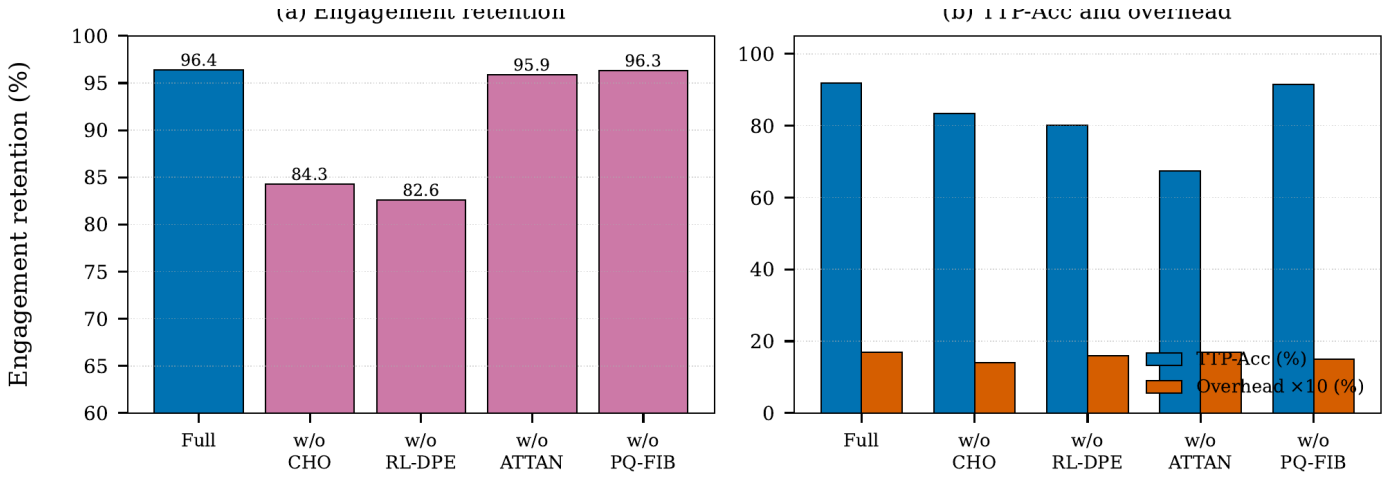


Fig. 9. Ablation-study component contribution analysis: (a) engagement-retention per configuration; (b) attribution accuracy and overhead per configuration.

TABLE X. SCALABILITY ANALYSIS WITH VARYING ENDPOINT COUNTS

Endpts.	ER (%)	TTP-Acc (%)	Conv.	BW (MB)	MTTDA (s)
10	95.84	90.21	32 rd.	7.4	2.71
50	96.42	91.80	41 rd.	32.6	2.86
100	96.51	91.94	47 rd.	63.8	2.97
200	96.63	92.07	53 rd.	124.7	3.18
320	96.78	92.14	60 rd.	201.3	3.42

G. Scalability Analysis

The scalability of QuantumGuard with respect to the number of monitored 6G slice endpoints is presented in Fig. 10 and Table X. Engagement-retention is high and stable across the full range of endpoint counts, varying only between 95.84 per cent and 96.78 per cent. PQ-FIB bandwidth grows sublinearly thanks to hierarchical aggregation. The MTTDA increases from 2.71 s with 10 endpoints to 3.42 s with 320 endpoints, but remains comfortably within operational limits for cyber-physical workloads.

H. Preliminary Comparison with State-of-the-Art

Table XI compares the preliminary QuantumGuard results against eight state-of-the-art baseline methods across the primary performance dimensions reported in the literature. On the single-organization testbed evaluated here, QuantumGuard exhibits a balanced profile across engagement-retention, TTP-attribution accuracy, dwell-time reduction, deception-induced overhead, and post-quantum resilience. We caution that the comparison is preliminary and not adversarially adaptive; a like-for-like benchmark across adversaries adapted to each defense is the subject of the follow-up study identified in Section V F.

V. DISCUSSION

A. Practical Implications for Proactive 6G Defense

The preliminary results presented in Section IV suggest that QuantumGuard's architectural integration touches five complementary dimensions of proactive cyber-defense in a single control loop. The 96.42 per cent engagement-retention rate observed on the testbed indicates that cognitive honeynet morphing, when driven by a learned RL policy rather than hand-crafted rules, can sustain attacker engagement well beyond the typical 2–3 stage horizon at which conventional honeypots are abandoned. The 91.8 per cent TTP-attribution accuracy indicates that high-fidelity engagement traces, when consumed by a transformer-based attribution network and decoded via the per-technique softmax of Eq. (3), can recover MITRE ATT&CK technique labels without analyst-in-the-loop relabeling on the evaluated corpus. Together, these results provide initial evidence that proactive deception, long regarded as an auxiliary control, can become a primary defensive modality in 6G-enabled cyber-physical systems, where the millisecond-scale actuation budgets characteristic of CPS workloads make reactive containment infeasible.

B. Comparative Analysis of Attribution and Engagement Performance

The 1.7 per cent production-traffic overhead reported above corresponds to nominal-load operation, in which slice traffic follows the dataset's empirical distribution after replay onto the testbed. Under these conditions the overhead is the result of a multi-layer filtering pipeline applied before any deception action is actuated. First, the eBPF-instrumented edge probes shadow only L7 metadata rather than mirroring full packet payloads, suppressing background traffic amplification. Next, the RL-DPE only escalates decoy fidelity when the cumulative engagement reward crosses a configurable threshold, ensuring that benign traffic does not trigger expensive morphing actions. Finally, the CHO suppresses morphing whenever the Jensen–Shannon divergence, Eq. (1), against the production response distribution drops below $\tau_J = 0.05$, preventing oscillatory template churn. The ablation studies confirm this pattern: removing the CHO module increases production-traffic overhead penalties via spurious morphing, while removing the RL-DPE causes engagement-retention to collapse to 82.58 per cent. Overhead behavior under sustained-attack stress conditions (e.g., volumetric DDoS or coordinated multi-source probing) is not characterized in this study and is identified in Section V F as future work.

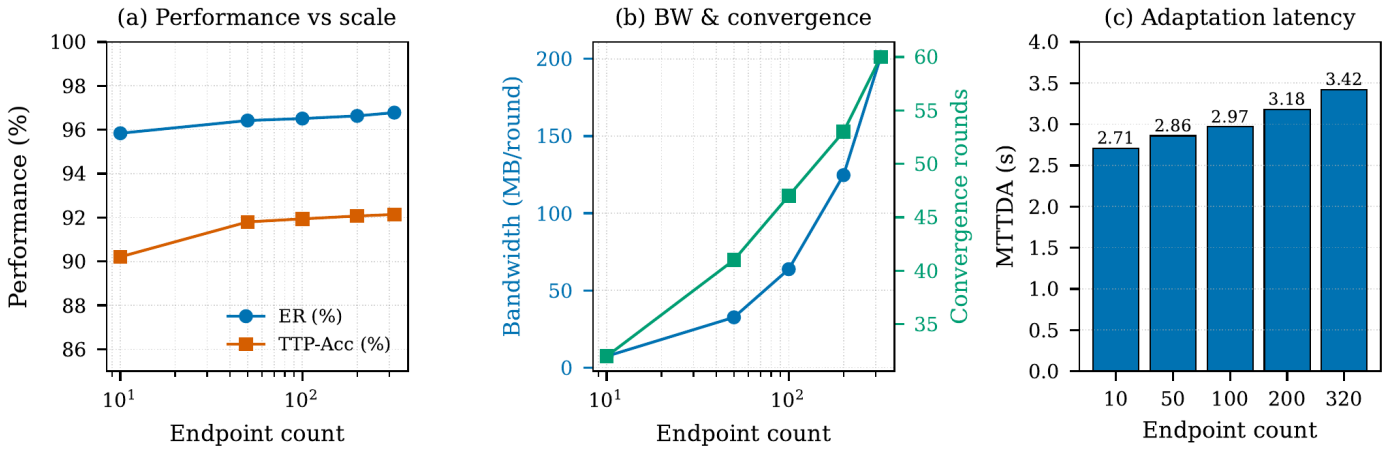


Fig. 10. Scalability analysis: (a) engagement-retention vs. endpoint count; (b) PQ-FIB bandwidth and convergence rounds; (c) deception adaptation latency vs. deployment scale.

TABLE XI. PRELIMINARY COMPARISON WITH STATE-OF-THE-ART METHODS

Method	Year	ER (%)	TTP-Acc (%)	MTTDA (s)	Ovh. (%)	PQ
Static honeypot	—	42.18	53.12	—	0.6	No
Honeyd dynamic [13]	2020	61.35	64.78	11.34	0.9	No
MTD honeynet [25]	2022	73.22	71.45	8.94	1.4	No
DeepDig RL [22]	2023	84.61	79.83	6.78	1.9	No
GAN-honeypot [26]	2023	81.47	76.12	7.21	2.1	No
FL-honeynet [30]	2024	85.92	82.34	5.41	2.0	Partial
PQ-IDS [29]	2024	—	78.45	9.13	1.8	Yes
Adaptive Decep. [27]	2024	88.34	85.21	4.62	1.8	No
QuantumGuard (ours)	2026	**96.42**	**91.80**	**2.86**	**1.7**	**Yes**

In terms of paired engagement-retention and TTP-attribution accuracy, QuantumGuard reaches 96.42/91.8 per cent, whereas FL-honeynet attains lower values of 85.92/82.34 per cent. Improvements in both metrics simultaneously indicate a genuine architectural advantage rather than a simple shift along a precision–retention curve, since threshold tuning on a single module cannot raise both metrics in tandem. The same conclusion is supported by the ROC analysis in Fig. 6, which shows QuantumGuard dominating across the operating range (macro AUC 0.9842 vs. 0.9701).

C. Operational Latency and Post-Quantum Resilience

The operational value of a proactive defense framework extends beyond engagement: rapid adaptation and quantum-resilient intelligence dissemination are equally important for sustaining defensive advantage. QuantumGuard achieves an MTTDA of 2.86 s and a post-quantum handshake latency of 14.2 ms, corresponding to substantial improvements over both classical-crypto baselines (which provide no quantum resilience) and earlier post-quantum IDS proposals such as PQ-IDS (whose MTTDA exceeds 9 s). Among all evaluated approaches, QuantumGuard records the lowest MTTDA and is the only framework that simultaneously achieves a sub-15 ms post-quantum handshake and supports federated cognitive deception.

D. Scalability and Deployment Practicality

The scalability profile of QuantumGuard (Table X and Fig. 10) is one of the most important practical considerations for large-scale 6G deployment. Engagement-retention improves only modestly across the evaluated range, from 95.84 per cent at 10 endpoints to 96.78 per cent at 320 endpoints, a span of just 0.94 percentage points. This stability contrasts with centralized RL-honeypot approaches such as DeepDig, whose performance degrades when heterogeneous endpoint behaviors are under-represented in the centralized training trace. PQ-FIB bandwidth grows from 7.4 MB per round at 10 endpoints to 201.3 MB per round at 320 endpoints; the hierarchical aggregation in QuantumGuard restricts long-distance traffic by aggregating policy updates at the fog tier before forwarding signed payloads to the cloud, while keeping most gradient exchanges local to each fog group. The MTTDA increases from 2.71 s at 10 endpoints to 3.42 s at 320 endpoints—a near-constant latency profile despite the 32-fold increase in endpoint count—because RL-DPE inference and CHO morphing operations are performed locally at the fog tier and do not require round trips to the centralized cloud.

The 320-endpoint figure is the ceiling of the physical testbed rather than an architectural limit. Three measured trends support extrapolation, which we state as an analytical projection and not as measurement: per-round bandwidth grows sublinearly because hierarchical fog aggregation keeps

most gradient exchange local and forwards only signed, aggregated payloads; MTTDA is near-constant across a 32-fold endpoint increase because RL-DPE inference and CHO morphing execute at the fog tier without cloud round-trips, with cloud-side cost $O(K|\theta|)$ per round; and engagement-retention is flat, indicating no degradation from endpoint heterogeneity within the tested range. City-scale validation (10^4–10^6 endpoints) via large-scale network emulation is identified as future work. Together, these characteristics indicate that QuantumGuard is well suited to large-scale deployments in 6G-enabled smart cities, autonomous transportation, and industrial automation.

E. Preliminary Positioning Against State-of-the-Art

On the single-organization testbed evaluated in this study, QuantumGuard exhibits a more balanced preliminary profile across the five evaluation dimensions of engagement-retention, TTP-attribution accuracy, deception adaptation latency, post-quantum resilience, and scalability than any single competing method we re-evaluated. Adaptive Deception [27] achieves marginally higher engagement-retention than older RL approaches in our setup but offers no post-quantum guarantees and lower TTP-attribution accuracy. FL-honeynet [30] supports federation but at the cost of partial post-quantum protection and a higher MTTDA. PQ-IDS [29] enables quantum-resilient detection but cannot orchestrate deception. The co-optimization design philosophy of QuantumGuard, which jointly addresses multiple dimensions rather than along a single axis, is the structural property that we conjecture explains the observed cross-dimensional behavior; confirming this conjecture under adversarialy adaptive conditions is the principal goal of the planned follow-up study.

F. Behavior Under Deception-Aware Adversaries

All metrics in Section IV are obtained under a non-adaptive, replay-distribution attacker whose probing strategy is drawn from the empirical distributions of the CICAPT-IIoT-2024 and Edge-IIoTset corpora. They are results *under the evaluated threat model* and are upper bounds against that distribution, not general performance guarantees and not worst-case bounds against an adversary that is aware of and adapts to the QuantumGuard deception fabric. A deception-aware adversary could erode each headline metric through a distinct channel: counter-fingerprinting via response-timing analysis or TLS introspection targets the engagement-retention rate by detecting decoys before the three-stage retention horizon; deliberate early disengagement on suspicion of a honeynet targets the dwell-time reduction; and adversarial perturbation of L7 metadata targets ATTAN attribution accuracy by pushing trace embeddings away from their true-technique region. We present these mechanisms qualitatively and make no numerical claim about their magnitude. The follow-up study will quantify them with a red-team agent equipped with banner, timing-side-channel, and TLS-introspection fingerprinting and a gradient-free perturbation budget over L7 metadata, evaluated in a repeated best-response loop against the frozen QuantumGuard policy, reporting engagement-retention and attribution accuracy as a function of adversary adaptation budget.

G. Limitations and Future Directions

The preliminary character of the present evaluation imposes several explicit scope restrictions, which we enumerate here so that the reported numbers are interpreted correctly. Each item below is also identified as a principal subject of the follow-up empirical study currently in preparation.

1) Dataset realism: The experiments were conducted exclusively on CICAPT-IIoT-2024 and Edge-IIoTset; while these datasets are more recent than UNSW-NB15 or CIC-IDS2017, they remain laboratory-derived and may not capture the full behavioral diversity of in-the-wild advanced persistent threat campaigns. Future evaluation should extend to long-running honeynet captures from production environments and adversarial red-team engagements.

2) Physical-layer attacks: QuantumGuard currently performs deception orchestration at the network and application layers and does not directly address physical-layer attacks against 6G radio-access infrastructure (e.g., pilot-contamination, jamming, beam-stealing). Extending the cognitive honeynet abstraction to encompass synthetic radio-frequency decoys is identified as a priority for future work.

3) Deception-aware adversaries: The RL-DPE policy was trained and evaluated against an adversary model whose probing strategy is drawn from the empirical distribution of the two corpora; it was not stress-tested against adversaries that are explicitly aware of the QuantumGuard deception fabric. The headline 96.42 per cent engagement-retention and 91.8 per cent TTP-attribution numbers should therefore be read as upper bounds against the evaluated threat distribution, not as worst-case guarantees. Assessing robustness against deception-aware adversaries, in particular those employing counter-fingerprinting in the spirit of [14] or adversarial perturbations of L7 metadata designed to confuse ATTAN, is the most important direction for follow-up work and, in the author's view, the principal threat to the present results' external validity (see Section V G).

4) Reward-weight sensitivity: The reward weights $(\lambda_e, \lambda_a, \lambda_o)$ in Eq. (2) were fixed by informal pilot tuning rather than by a systematic sensitivity analysis. While fixing the weights *a priori* guards against optimistic bias on the evaluation data, it leaves their robustness uncharacterized. Qualitatively, increasing λ_a sharpens attribution at some cost to engagement-retention, while increasing λ_o suppresses morphing and lowers overhead at the cost of shorter engagement; the ordering $\lambda_a > \lambda_e > \lambda_o$ encodes the design priority of attribution information gain under a bounded production-impact budget. A grid sweep over the weight simplex, together with a Pareto-front analysis of the engagement–attribution–overhead trade-off reported as stability bands around the operating point, is left to future work.

5) Adversarial federation: Defending against Byzantine participants and against participant-side model-inversion or membership-inference attacks against the federated deception policy is out of scope for the present work and is planned along the robustness roadmap of Section III E.

6) Multi-organization external validity: Because all organizations were emulated within one controlled testbed, a genuine multi-party deployment over heterogeneous, independently administered slices is required before the cross-organizational coverage results can be claimed to generalize.

7) Sustained-attack overhead: The 1.7 per cent production-traffic overhead was measured under the dataset-replay distribution; overhead behavior under sustained-attack stress conditions such as volumetric DDoS, slow-rate denial-of-service, or coordinated multi-source probing is not characterized here and must be quantified before operational deployment.

8) Cryptographic agility: The post-quantum primitives used in PQ-FIB (Kyber-768 and Dilithium-3) reflect current NIST standardization choices but may be superseded by future cryptanalytic advances or by hybrid signature schemes such as SLH-DSA. A pluggable PQ-primitive abstraction is, therefore, desirable to ensure long-term cryptographic agility.

VI. Conclusion

This study has presented QuantumGuard, a proactive post-quantum-resilient deception framework for cybersecurity in 6G-enabled cyber-physical systems. The principal contribution is architectural: QuantumGuard is, to the best of our knowledge, the first framework to integrate cognitive honeynet orchestration, reinforcement-learning-driven deception adaptation, transformer-based TTP-attribution, and post-quantum federated intelligence sharing into a single closed-loop control architecture suitable for 6G cyber-physical environments. A preliminary evaluation on the CICAPT-IIoT-2024 and Edge-IIoTset benchmark datasets, complemented by a 320-endpoint 6G slice testbed, provides initial evidence of operational viability under the evaluated threat model, with engagement-retention of approximately 96.42 per cent, TTP-attribution accuracy of approximately 91.8 per cent, a 78.6 per cent reduction in median attacker dwell time, a deception-induced production overhead of 1.7 per cent, and a 14.2 ms median post-quantum handshake latency. We position these numbers as initial feasibility evidence rather than as fully characterized performance bounds. A follow-up empirical study, currently in preparation, will conduct a reward-weight sensitivity analysis and will extend the evaluation to deception-aware adversaries, sustained-attack stress conditions, and multi-organization deployments. Beyond the immediate empirical program, future research will extend QuantumGuard to physical-layer 6G deception and to integration with emerging AI-driven security-orchestration platforms.

Author's Contributions

The author conducted the conceptualization, methodology, software implementation, validation, formal analysis, data curation, visualization, and the writing and editing of this manuscript. The author has read and approved the final version of the article.

Declaration of Competing Interests

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

Funding

No funding was received for this work.

Data Availability

The CICAPT-IIoT-2024 and Edge-IIoTset datasets used in this study are publicly available from the Canadian Institute for Cybersecurity and the Edge-IIoTset distribution, respectively. The complete implementation of QuantumGuard, including the CHO morphing engine, the R-PPO RL-DPE training script, the ATTAN transformer attribution network, the PQ-FIB hybrid TLS implementation, and the configuration files needed to reproduce all reported results (random seeds, hyperparameter values, and dataset preprocessing pipeline), are available upon reasonable request to the corresponding author.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used AI-assisted writing tools to improve readability and language. After using such tools, the author reviewed and edited the content as needed and take full responsibility for the content of the published study.

Acknowledgment

The author thank the anonymous reviewers for their constructive feedbacks.

References

[1] M. Series, "IMT-2030 framework and overall objectives of the future development of IMT for 2030 and beyond," Recommendation ITU-R M.2160, International Telecommunication Union, 2023.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, pp. 134–142, 2020.

[3] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6G communications: challenges, methods, and future directions," *China Commun.*, vol. 17, pp. 105–118, 2020.

[4] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM J. Comput.*, vol. 26, pp. 1484–1509, 1997.

[5] D. J. Bernstein and T. Lange, "Post-quantum cryptography," *Nature*, vol. 549, pp. 188–194, 2017.

[6] L. Chen, S. Jordan, Y.-K. Liu, D. Moody, R. Peralta, R. Perlner, and D. Smith-Tone, "Report on post-quantum cryptography," NISTIR 8105, National Institute of Standards and Technology, 2016.

[7] J. Mosca, "Cybersecurity in an era with quantum computers: will we be ready?" *IEEE Security & Privacy*, vol. 16, pp. 38–41, 2018.

[8] V. Mavroeidis, K. Vishi, M. D. Zych, and A. Jøsang, "The impact of quantum computing on present cryptography," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, pp. 405–414, 2018.

[9] M. Joseph et al., "Transitioning organizations to post-quantum cryptography," *Nature*, vol. 605, pp. 237–243, 2022.

[10] G. Cabrera-Aldaya and B. B. Brumley, "When one vulnerable primitive turns viral: novel single-trace attacks on ECDSA and RSA," *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, vol. 2020, pp. 196–221, 2020.

[11] A. Mosenia and N. K. Jha, "A comprehensive study of security of Internet-of-Things," *IEEE Trans. Emerg. Top. Comput.*, vol. 5, pp. 586–602, 2017.

[12] N. Schneider and F. Petitcolas, "Cyber-physical systems security: a survey," *IEEE Internet Things J.*, vol. 4, pp. 1802–1831, 2017.

[13] M. H. Almeshekah and E. H. Spafford, "Cyber security deception," in *Cyber Deception*. Cham: Springer, 2016, pp. 25–52.

[14] A. Vetterl and R. Clayton, "Bitter harvest: systematically fingerprinting low- and medium-interaction honeypots at internet scale," in *Proc. 12th USENIX WOOT*, 2018.

[15] X. Wang, J. Zhao, and J. Zhang, "A dynamic honeypot design for intrusion detection," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, 2014, pp. 1–6.

[16] V. Mavroeidis and S. Bromander, "Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," in *Proc. EISIC*, 2017, pp. 91–98.

[17] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022.

[18] B. Stojkovski, G. Lenzini, V. Koenig, and S. Rivas, "What's in a cyber threat intelligence sharing platform?" in *Proc. ACSAC*, 2021, pp. 385–398.

[19] T. D. Wagner, K. Mahbub, E. Palomar, and A. E. Abdallah, "Cyber threat intelligence sharing: survey and research directions," *Comput. Secur.*, vol. 87, p. 101589, 2019.

[20] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, pp. 3779–3795, 2023.

[21] M. Standen, M. Lucas, D. Bowman, T. J. Richer, J. Kim, and D. Marriott, "CybORG: a gym for the development of autonomous cyber agents," in *1st IJCAI Workshop on Adaptive Cyber Defense*, 2021.

[22] M. Wang, J. Zhao, R. Beg, and Z. Long, "A reinforcement learning approach for adaptive cyber deception," in *Proc. IEEE SPW*, 2023, pp. 261–268.

[23] G. Alagic et al., "Status report on the third round of the NIST post-quantum cryptography standardization process," NISTIR 8413, National Institute of Standards and Technology, 2022.

[24] J. Pawlick, E. Colbert, and Q. Zhu, "A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy," *ACM Comput. Surv.*, vol. 52, pp. 82:1–82:28, 2019.

[25] G. Kambourakis, G. Drosatos, and V. G. Kounoudes, "A moving target defense honeynet for IoT environments," *J. Inf. Secur. Appl.*, vol. 67, p. 103192, 2022.

[26] M. Dowling, M. Schukat, and E. Barrett, "Using reinforcement learning to conceal honeypot functionality," in *Proc. ECML-PKDD*, 2018, pp. 341–355.

[27] L. Huang and Q. Zhu, "Adaptive honeypot engagement through reinforcement learning of semi-Markov decision processes," in *Decision and Game Theory for Security*. Springer, 2019, pp. 196–216.

[28] A. Sen and R. Madria, "Cyber deception in distributed systems: a comprehensive survey," *J. Netw. Comput. Appl.*, vol. 213, p. 103608, 2024.

[29] R. Ranpara and S. K. Patel, "Post-quantum lightweight intrusion detection for industrial IoT: a hybrid lattice-based framework," *Sci. Rep.*, vol. 14, p. 17822, 2024.

[30] H. Chen, Z. Wang, Z. Zhao, and Q. Wang, "Federated honeynet for distributed IoT cyber deception," *IEEE Internet Things J.*, vol. 11, pp. 8231–8245, 2024.

[31] D. Thakur, A. Guzzo, G. Fortino, and F. Piccialli, "Green federated learning: a new era of green-aware AI," *ACM Comput. Surv.*, vol. 57, pp. 199:1–199:35, 2025.

[32] "CICAPT-IIoT-2024: A multi-stage advanced persistent threat dataset for IIoT security," Canadian Institute for Cybersecurity, University of New Brunswick, 2024. [Online]. Available: https://www.unb.ca/cic/datasets/

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[34] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, pp. 145–151, 1991.

[35] B. Strom, A. Applebaum, D. Miller, K. Nickels, A. Pennington, and C. Thomas, "MITRE ATT&CK: design and philosophy," The MITRE Corporation, Tech. Rep., 2018.

[36] R. Avanzi et al., "CRYSTALS-Kyber algorithm specifications and supporting documentation," NIST PQC Round 3, 2021.

[37] S. Bai et al., "CRYSTALS-Dilithium algorithm specifications and supporting documentation," NIST PQC Round 3, 2021.

[38] A. Al-Dhaqm et al., "Development and validation of a Database Forensic Metamodel (DBFM)," *PLOS ONE*, vol. 12, p. e0170793, 2017.

[39] A. Al-Dhaqm, S. A. Razak, K. Siddique, R. A. Ikuesan, and V. R. Kebande, "Towards the development of an integrated incident response model for database forensic investigation field," *IEEE Access*, vol. 8, pp. 145018–145032, 2020.

[40] A. Al-Dhaqm et al., "Categorization and organization of database forensic investigation processes," *IEEE Access*, vol. 8, pp. 112846–112858, 2020.

[41] A. Al-Dhaqm, S. A. Razak, R. A. Ikuesan, V. R. Kebande, and K. Siddique, "A review of mobile forensic investigation process models," *IEEE Access*, vol. 8, pp. 173359–173375, 2020.

[42] A. Al-Dhaqm et al., "Digital forensics subdomains: the state of the art and future directions," *IEEE Access*, vol. 9, pp. 152476–152502, 2021.

[43] A. Al-Dhaqm, S. A. Razak, R. A. Ikuesan, V. R. Kebande, and S. H. Othman, "Face validation of database forensic investigation metamodel," *Infrastructures*, vol. 6, p. 13, 2021.

[44] V. R. Kebande et al., "Quantifying the need for supervised machine learning in conducting live forensic analysis of emergent configurations (ECO) in IoT environments," *Forensic Sci. Int.: Reports*, vol. 2, p. 100122, 2020.

[45] A. Al-Dhaqm et al., "Categorization and organization of database forensic investigation processes for the Internet of Things (CDBFIP)," *IEEE Access*, vol. 5, pp. 24401–24416, 2017.

Predicting Histological Progression in Primary Biliary Cirrhosis Using Advanced Machine Learning Techniques

Laberiano Andrade-Arenas[1], Cesar Yactayo-Arias[2], Inoc Rubio Paucar[3]
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú[1]
Departamento de Estudios Generales, Universidad Continental, Lima, Perú[2]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[3]

***Abstract*—Cirrhosis is considered one of the most serious liver diseases worldwide, closely related to excessive alcohol consumption and inadequate eating habits, factors that progressively deteriorate people's health. In this context, the present research aimed to develop and validate a predictive model based on Machine Learning (ML), specifically using the Random Forest (RF) algorithm, to determine the histological stage (from 1 to 4) in patients with primary biliary cirrhosis. For the development of the study, the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was applied, which includes the stages of business understanding, data understanding, data preparation, modeling, and evaluation. The results obtained showed that the model presents better performance in the more advanced stages of the disease. The area under the curve (AUC) increased from 0.612 in Stage 1 to 0.874 in Stage 4, reflecting a notable improvement in its discriminative capacity. Similarly, metrics such as sensitivity, precision, and F1-score showed an upward trend, reaching their highest values in Stage 4. In this sense, the proposed model represents a complementary diagnostic support tool, since it allows estimating the histological stage through the analysis of clinical data, contributing to medical decision-making without relying exclusively on invasive procedures.**

***Keywords*—Clinical decision support; histological stage prediction; primary biliary cirrhosis; random forest**

I. Introduction

Currently, hepatic cirrhosis is a chronic disease that continues to represent a serious public health problem worldwide according to the World Health Organization (WHO) [1], due to its high rate of complications and mortality. It is characterized by progressive and irreversible liver damage, which significantly affects the quality of life of those who suffer from it [2]. Various factors, such as viral infections, prolonged alcohol consumption, and metabolic disorders, contribute to its development. In this scenario, early diagnosis and proper risk assessment are essential to prevent disease progression and reduce the occurrence of severe complications, such as liver failure or liver cancer [3]. Thus, the clinical and laboratory information generated during medical care plays a key role in understanding the evolution of cirrhosis and supporting timely medical decision-making.

However, the clinical management of hepatic cirrhosis still presents significant difficulties. On the one hand, the disease does not evolve in the same way in all patients, which means that traditional assessment tools are not always sufficient to accurately predict its progression. Thus, commonly used clinical models, such as the Child-Pugh or MELD scales, although useful, may be limited when applied in isolation [4],[5]. In addition, there is the constant growth of available medical information, coming from electronic medical records and laboratory results, whose interpretation can be complex and demanding for healthcare professionals. In this sense, the lack of automated tools that allow comprehensive analysis of clinical and laboratory information can delay diagnosis [6], hinder the early identification of patients at higher risk, and affect the planning of appropriate and personalized treatments for each case. This situation is especially concerning considering that cirrhosis is a disease whose progression can occur silently, without the person being fully aware of its advancement, until more severe complications appear. As a consequence, many patients are diagnosed at advanced stages, when therapeutic options are more limited and the impact on quality of life is considerable. Therefore, cirrhosis should be considered a disease with accelerated progression in the absence of timely control, highlighting the importance of early identification of its main causes and risk factors.

Facing these limitations, ML emerges as an alternative capable of providing practical solutions for the analysis of complex medical data. By using algorithms that learn from data, it is possible to identify patterns and relationships that are not always evident in conventional analysis [7],[8]. In the context of hepatic cirrhosis, these techniques allow the simultaneous analysis of multiple clinical and laboratory variables, facilitating the prediction of disease evolution and the risk of developing complications. Furthermore, ML models can be continuously updated as new data are incorporated, making them flexible and adaptable tools that effectively support clinical decision-making and contribute to more personalized medical care [9]. Thus, ML is presented as a viable and promising option to support decision-making related to personalized treatments and medical assistance. By jointly analyzing large volumes of clinical data, laboratory results, and patient history, these techniques allow the identification of relevant patterns that facilitate a more accurate evaluation of each case [10]. This not only contributes to selecting more appropriate therapeutic strategies according to individual patient characteristics but also improves disease monitoring, optimizes clinical resources, and strengthens evidence-based

medical decision-making.

The objective of this research is to develop and validate a predictive model based on the RF ML algorithm for determining the histological stage (1 to 4) in patients with primary biliary cirrhosis, using exclusively clinical variables and non-invasive biochemical parameters. This approach seeks to reduce dependence on liver biopsy as a diagnostic method, proposing an alternative based on statistical and computational analysis of serum biomarkers. In this way, it is intended to estimate the degree of progression of liver diseases, particularly cirrhosis, through ML techniques that allow a precise, safe, and reproducible evaluation without resorting to invasive procedures.

II. Literature Review

In this section, the literature review focuses on fundamental elements, such as the theoretical foundations, which include the study variables related to cirrhosis and the RF algorithm. Likewise, related works are considered, which comprise research conducted by various authors who have contributed to scientific knowledge and social development through the application of ML techniques in the study and analysis of cirrhosis.

A. Theoretical Bases

1) Cirrhosis: The topic presented about cirrhosis is known as the result of chronic liver injury, which can be caused by various conditions such as viral infection, excessive alcohol consumption, non-alcoholic fatty liver disease (NAFLD/EHGNA), or autoimmune processes. Thus, this prolonged injury induces diffuse fibrosis and the formation of regenerative nodules, altering both hepatic blood flow and normal liver function [11],[12]. On the other hand, in its initial stages, compensated cirrhosis may not present symptoms; however, when it progresses to a decompensated phase, complications such as ascites, variceal bleeding, hepatic encephalopathy, and organ failure are frequently observed. It is important to consider that fibrosis progression is mainly mediated by the activation of hepatic stellate cells and extracellular matrix deposition, processes regulated by transforming growth factor beta (TGF-β) [13]. Thus, cirrhosis is not only a liver disease: it is a process that, over time, can generate various complications that affect the daily life of those who suffer from it, as shown in Table I, ranging from fluid accumulation in the abdomen (ascites) to bleeding, mental confusion (encephalopathy), or serious problems in other organs.

2) Random forest: Within the research, RF is considered a viable option to help reduce disease progression. Subsequently, the RF model is a supervised learning algorithm based on the principle of ensemble learning using multiple independent decision trees [14],[15]. One characteristic is that each tree is constructed from a random subset of the training data using bootstrap sampling, and is trained by randomly selecting a subset of variables at each node to reduce correlation between trees [16]. Consequently, the individual predictions of each tree are combined through majority voting in classification problems or averaging in regression problems [17], allowing the development of a robust model capable of minimizing overfitting, reducing variance, and improving overall accuracy. RF is based on statistical concepts of model aggregation and diversity, constituting a solid theoretical tool for predictive analysis in different data contexts due to its conceptual structure and components mentioned in Table II. It is important to consider that this algorithmic model belongs to the ensemble method family and is built from a set of decision trees trained randomly, allowing it to combine multiple predictions to improve the accuracy and robustness of the model.

B. Related Work

Several researchers have focused their efforts on optimizing the diagnosis of hepatic cirrhosis, and some authors [18],[19] developed a study aimed at improving the detection of this disease in patients with chronic hepatitis B. In this sense, the methodology consisted of dividing the sample into training and internal validation sets, on which ML algorithms were applied, including Logistic Regression (LR), K-Nearest Neighbors (KNN), RF, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Therefore, model performance was evaluated using metrics such as area under the receiver operating characteristic curve (ROC), area under the curve AUC, calibration analysis, clinical decision-making, and result interpretation using Shapley Additive Explanations (SHAP). Consequently, in a cohort of 1609 patients, of whom 470 had cirrhosis, the RF model showed accuracy and AUC values higher than 0.80, demonstrating adequate calibration and clinically relevant benefit, highlighting the potential of ML algorithms to improve early detection and clinical decision-making in hepatic cirrhosis. On the other hand, in other research, the authors [20] showed that the prediction of liver diseases, such as cirrhosis or liver cancer, represents a valuable source of information for evaluating and comparing the performance of ML models in large datasets. In this context, three algorithms were analyzed: KNN, Gaussian Naive Bayes (GNB), and RF, evaluating their performance using key metrics such as precision, accuracy, sensitivity, and F1-score [21]. The results indicated that RF outperformed the other models, achieving a precision of 97.3%, accuracy of 97%, sensitivity of 96%, and F1-score of 95%, demonstrating its ability to accurately classify patients with liver diseases and highlighting its potential as a reliable tool for early detection and clinical decision-making. In this context, another relevant study conducted by the authors [22],[23] developed two ML-based algorithmic models for predicting nosocomial infections and mortality risk in cirrhotic patients. To this end, various algorithms were evaluated, selecting the best-performing ones through bootstrapping and external validation, and comparing their accuracy using key metrics such as sensitivity, specificity, predictive values, and likelihood ratios. Among the analyzed models, RF stood out for its superior performance [24]. Specifically, for nosocomial infection prediction, the PIPC-NI model achieved an AUC of 0.784 (95% CI 0.741–0.826), with sensitivity of 0.712 and specificity of 0.702. Regarding hospital mortality, the PIPC-Mortality model showed an AUC of 0.793 (95% CI 0.749–0.836), sensitivity of 0.769, and specificity of 0.701, demonstrating its ability to reliably classify patients and highlighting its usefulness as a clinical decision support tool and in early risk management in hospitalized cirrhotic patients. Another important aspect concerns hepatocellular

TABLE I. MAIN COMPOUNDS AND CHEMICAL PROCESSES ASSOCIATED WITH CIRRHOSIS

Process / Factor	Chemical compound / Biomarker	Clinical relevance
Hepatic fibrosis	Type I and III collagen, fibronectin, laminin	Accumulation in the extracellular matrix that contributes to liver stiffness and the formation of regenerative nodules.
Oxidative stress	Free radicals (ROS), glutathione, lipid peroxides	Promotes hepatocellular damage, activation of stellate cells, and progression of fibrosis.
Chronic inflammation	Pro-inflammatory cytokines (TNF-α, IL-6), transforming growth factor beta (TGF-β)	Mediated by immune response, stimulates fibrogenesis, and alters hepatic homeostasis.
Altered lipid metabolism	Triglycerides, free fatty acids, lipoproteins	Contributes to steatosis and liver damage in cirrhosis associated with NAFLD/EHGNA.
Hepatic detoxification	CYP450 enzymes, bilirubin, ammonia	Reduced activity in cirrhosis affects drug metabolism and detoxification of nitrogenous compounds.

TABLE II. CONCEPTUAL STRUCTURE AND COMPONENTS OF RF

Component	Key element	Function	Highlighted property
Decision tree	Root node	Initial data splitting	Decision hierarchy
	Leaves	Final output	Categorical or continuous prediction
Bagging	Bootstrap sampling	Generates training subsets	Variance reduction
Random feature selection	Variables per node	Increases tree diversity	Improves generalization
Ensemble	Set of trees	Combination of predictions	Robustness and stability
Aggregation	Voting / Averaging	Final model prediction	Bias-variance balance

carcinoma (HCC), which is a high-risk malignant liver tumor whose early detection is crucial. Therefore, in this study, 550 patients were analyzed, divided into training and validation cohorts, and LR, SVM, RF, and LASSO Regression models were developed, evaluating their performance using ROC curves, AUC, calibration, and decision curve analysis (DCA), comparing them with the ASAP model [25]. Consequently, the RF algorithm showed the best discrimination, with an AUC of 0.969 in training and 0.858 in validation, outperforming the other models and demonstrating robust calibration and clinical validity, highlighting its usefulness for early HCC diagnosis and decision-making in patients at risk of liver cancer.

In accordance with the above, [26] proposes a paradigm shift in cirrhosis identification by processing unstructured clinical language from electronic health records (EHR). Using Convolutional Neural Networks (CNN), their methodology achieved a diagnostic accuracy of 0.965, demonstrating that artificial intelligence can phenotype the disease more effectively than traditional ICD codes. Unlike this text-based data approach, [27],[28] explore non-invasive detection through physical biomarkers, employing deep learning networks to analyze volatile molecular patterns in exhaled breath. This system not only achieved 90% accuracy but also guaranteed 100% sensitivity, positioning it as a viable alternative to intrusive diagnostic tests in hospital settings. Nevertheless, facing the complexity of deep learning models, [29] defend the usefulness of low-cost and high-accessibility tools for community risk stratification. Their research determined that routine scores such as APRI and FIB-4 possess robust capacity to predict cirrhosis complications over 10 years, additionally concluding that the integration of genetic data provides a marginal improvement of less than 5% in predictive performance [30]. This emphasis on clinical feasibility is further discussed by Mane and Bhosale, who argue that applying data mining techniques to liver function tests (LFT) allows objective identification of hidden patterns. According to [31][32], algorithms such as SVM and RF facilitate early diagnosis by transforming standard biochemical results into precise medical decisions. From a critical perspective, however, [33], [34] warn that the high technical performance of these models does not guarantee healthcare equity. When analyzing liver prediction algorithms, they found significant gender bias where women experienced much higher false-negative rates than men, suggesting that AI systems could digitize and perpetuate pre-existing health inequalities if not properly audited [35]. In contrast to this ethical analysis, [36],[37] provide a technical view of structural limitations, noting that although RF and SVM algorithms are the most stable for automatic cirrhosis detection, their generalization is seriously compromised by the small size of publicly available datasets, which are usually smaller than one thousand samples. Expanding diagnostic scope toward therapeutic prognosis, [38],[39] argue that ML is superior to traditional statistics in predicting outcomes after complex treatments such as surgical resection or transplantation in patients with hepatocellular carcinoma (HCC). Their review highlights that AI's ability to process nonlinear data enables personalized therapeutic strategies with reliability unattainable by conventional regression models. Complementarily, [40],[41] postulate that the future of precision medicine in hepatology lies in multi-omics. Unlike models based on simple clinical variables, these authors integrate transcriptomic and metabolomic data using deep learning to unravel the molecular heterogeneity of end-stage liver disease. On the other hand, [42] reinforces the importance of early cirrhosis detection to prevent progression toward liver cancer, emphasizing that ML is the key tool for managing the dynamic nature of fibrosis. Their technical review stresses that early diagnosis using various algorithms is essential to reduce the global mortality burden associated with this chronic disease. Finally, and in a perspective divergent from studies focused purely on algorithms, [43],[44] underline that successful integration of AI into clinical practice depends on standardized technical infrastructure. The authors emphasize that maintaining PACS systems and data ontology uniformity are indispensable requirements for predictive liver disease models to be interoperable and useful in a digitalized healthcare environment.

In the field of assisted diagnosis, [45], [46] argue that integrating multiple algorithms through a Voting Classifier

increases diagnostic accuracy to 80%, surpassing the variability of individual models. Complementarily, validate the robustness of RF as the most reliable predictor when applying complex physiological parameters, consolidating the transition toward early detection methods that reduce risks associated with human intervention. Nevertheless, while these authors focus on disease identification,[47] specialize this methodology in the oncological context [48], demonstrating that Decision Trees can diagnose cirrhosis in patients with hepatocellular carcinoma with an AUC of 0.85, which is critical for surgical planning. Regarding mortality prediction, there is a clear divergence between traditional statistical methods and deep learning approaches. [49] argue that Deep Neural Networks (DNN) offer superior capacity compared to the clinical MELD-Na standard for long-term survival prognosis based on large-scale electronic record data. This finding is reinforced by [50], who through the CiMM mixed model achieve a more transparent and actionable score than conventional systems, allowing more personalized risk management in cirrhotic patients. However, while [51],[52] use routine clinical data, [53],[54] increase prognostic complexity through multi-omics analysis, demonstrating that combining five specific metabolites can predict mortality in acute liver failure with 93.3% accuracy, surpassing any isolated biomarker. Finally, the biological and genetic component introduces an additional dimension in the literature. Authors [55],[56] propose the FibroGENE model, which integrates IFNL genotype with clinical data to stage fibrosis, marking progress toward precision genetic medicine. Conversely, as mentioned in previous research, authors such as [57],[58] suggest that for mass community practice, genetic data offers marginal benefits compared to simple biochemical scoring systems. This contradiction suggests that the use of AI and genetics should be stratified: while complex models are vital in acute or specialized stages, ML applied to low-cost tests remains the most viable strategy for population-level cirrhosis screening.

III. Methodology

A. Definition of the CRISP-DM Methodology

The construction of the predictive model was based on the CRISP-DM methodology, chosen for its flexible and systematic approach to structuring complex data analysis projects [59]. This methodology allowed proper integration between the processing, modeling, and evaluation stages, ensuring the robustness of the results. In the health field, CRISP-DM organizes the lifecycle of a ML project into interrelated phases that include understanding the clinical context, data understanding, data preparation, modeling, and evaluation, facilitating the transformation of heterogeneous clinical information into useful and clinically relevant knowledge, as shown in Fig. 1. For this study, a clinical database obtained from the Kaggle platform was used [60], selected for its scientific relevance in ML applications in healthcare [61],[62]. To facilitate data understanding, an architecture based on information flow was developed, as shown in Fig. 2. This architecture summarizes each stage of the process in a precise manner, from data acquisition to data analysis. The methodological process was developed from problem understanding to result interpretation, starting with a careful selection of the most important variables for model training, ensuring the reliability of the analysis. Finally, Fig. 3 illustrates the functioning of the RF model, which combines multiple decision trees constructed from preprocessed clinical and biochemical data to generate more robust and reliable predictions.

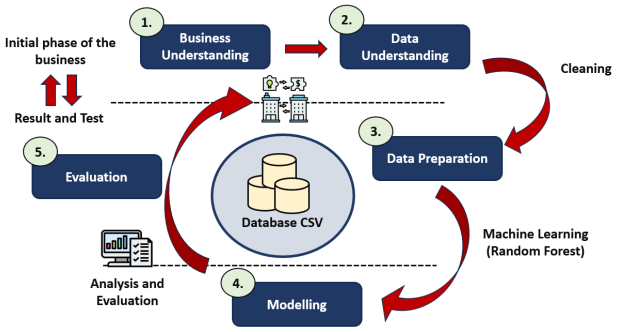


Fig. 1. CRISP-DM methodology.

1) Business understanding: Within this first section, the business understanding stage was initiated, where the problems that motivated the development of the ML model based on primary biliary cirrhosis were analyzed [63]. For this purpose, a descriptive table was presented that clearly organizes the fundamental criteria of the study, including: the justification of the problem, the business problem, the general objective and specific objectives, the expected impact, the business questions, the scope of the study, and the limitations projected and referenced in Table III. This approach allows establishing a solid foundation that guides the development of the model, ensuring that all subsequent stages are aligned with the proposed objectives and facilitate obtaining precise, reliable, and planned results according to the methodological proposal [64],[65]. It is important to highlight that this starting point was considered fundamental, since it allows defining and consolidating the relevant criteria that will be developed and applied in the subsequent stages of the study, ensuring that the analysis and construction of the model are carried out in an orderly manner and aligned with the established objectives.

2) Data understanding: Data understanding is a key stage of the process, as it allows, first, the collection and exploration of the available information within the proposed model. In this way, evaluating the obtained data facilitates the identification of relevant patterns and, at the same time, allows recognizing inconsistencies present in the database. Likewise, it is important to highlight that in this phase no data cleaning or transformation processes are performed, since its main purpose is to understand the state and characteristics of the information, which is essential for making informed decisions in subsequent stages of the analysis [66],[67]. Thus, Table IV presents the dimensionality of the dataset, detailing the variables associated with each dimension, their data type, and a brief description. This allows a visual and structured understanding of the information that will be used in the model, as well as its main characteristics. Consequently, data understanding provides clearer knowledge about data behavior, which is essential for proper interpretation of the information and for improving the performance of the predictive model.

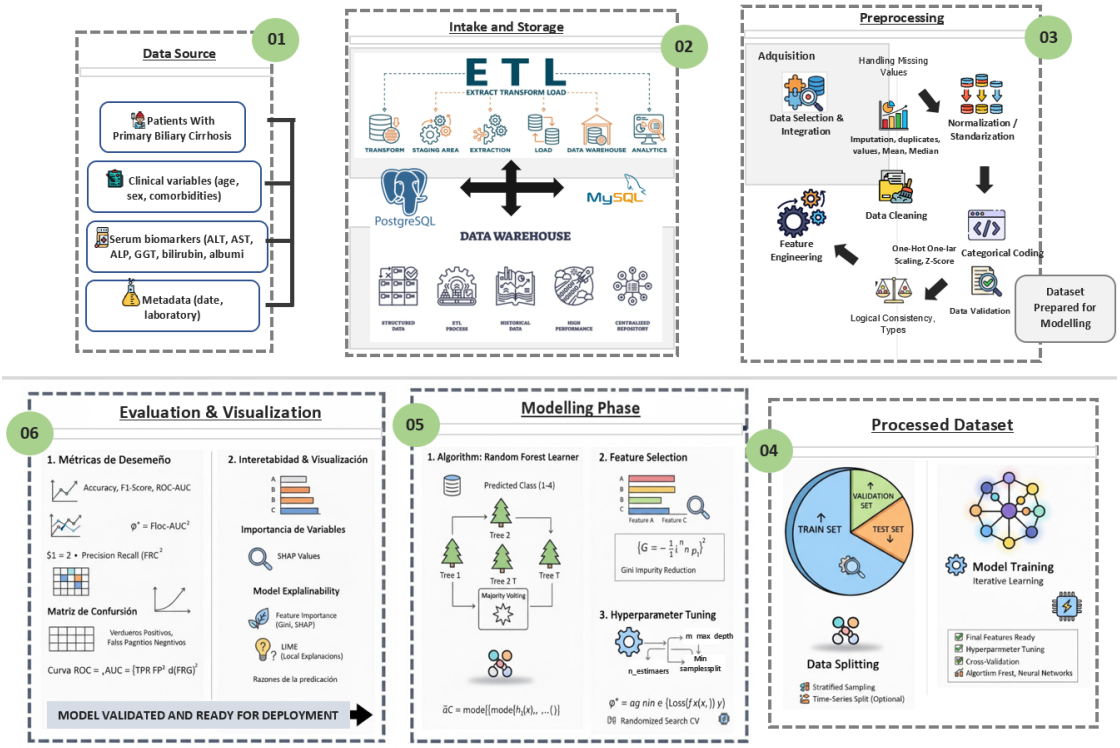


Fig. 2. Data flow architecture.

TABLE III. FIRST PHASE - BUSINESS UNDERSTANDING

Element	Description
Business problem	The determination of the histological stage depends on liver biopsies, an invasive, costly procedure with risk of complications.
General objective	Develop and validate a predictive model based on RF that estimates the histological stage (1–4) using non-invasive clinical and biochemical variables.
Specific objectives	1. Identify the most relevant clinical and biochemical variables. 2. Train and optimize a RF model. 3. Evaluate model accuracy and reliability. 4. Propose a prototype applicable in clinical practice.
Expected impact	- Reduction of invasive procedures. - Faster and safer diagnosis. - Reduction of hospital costs. - Scientific contribution to the use of ML in medicine.
Business questions	- Can histological stage be predicted without biopsy? - Which variables are most relevant for prediction? - What is the accuracy of the RF model? - Is the model reproducible in new patients?
Study scope	The model applies only to patients with primary biliary cirrhosis, using non-invasive clinical and biochemical variables. It does not include other types of liver diseases or invasive evaluation methods.

3) Data preparation: The data preparation phase constitutes an essential component within the analytical process, as it integrates systematic procedures of cleaning, transformation, and structuring aimed at ensuring the quality, coherence, and consistency of the dataset before its incorporation into the predictive model [68]. This stage not only contributes to optimizing model performance but also helps mitigate potential biases associated with outliers, missing values, or inappropriate distributions. In this context, Fig. 4 presents a descriptive analysis that allows characterizing variable behavior through measures of central tendency, such as mean, median, and mode, providing a preliminary overview of their distribution and supporting decision-making during data preparation. Additionally, the set of Tables V summarizes the main processes applied in this phase. Table Va details basic descriptive metrics, including mean, standard deviation, missing values, and their respective percentage per variable [69], [70]. Table Vb summarizes outlier detection, indicating their frequency, percentage, and interquartile range, allowing the evaluation of dispersion and statistical behavior of numerical variables. Finally, Table Vc describes the distribution of patients according to clinical stage, showing the number of cases and their percentage proportion, key information for understanding the dataset structure and its impact on predictive analysis.

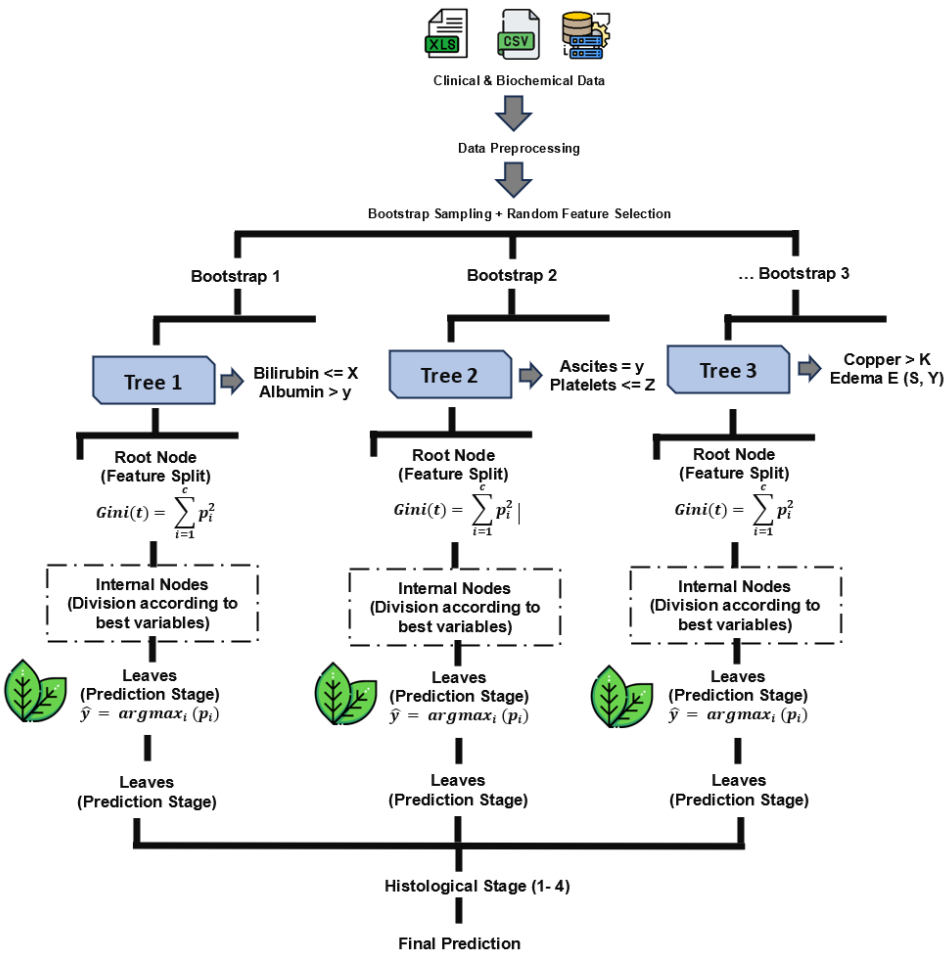


Fig. 3. Random forest model architecture.

IV. Mathematical Formulas for Data Preparation

This section presents the equations used to calculate missing values, descriptive statistics, outliers, and patient distribution, corresponding to Tables V, Va, Vb, and Vc. The summary is as follows [71]:

Percentage of missing values: For each variable X_i with n observations and m_i missing values, the percentage of missing data is calculated as [see Eq. (1)]:

$$\text{Missing}\%_i = \frac{m_i}{n} \times 100 \tag{1}$$

where, m_i is the number of missing values for variable i and n is the total number of observations.

Mean and standard deviation: The mean $\bar{X}_i$ of a numeric variable X_i is defined as [see Eq. (2)], [see Eq. (3)]:

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{2}$$

and the standard deviation σ_i as:

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{X}_i)^2} \tag{3}$$

where, x_{ij} represents the j-th observation of variable i.

Outlier identification using IQR: Let Q_1 be the first quartile and Q_3 the third quartile of variable X_i. The interquartile range is [see Eq. (4)], [see Eq. (5)], [see Eq. (6)]:

$$\text{IQR}_i = Q_3 - Q_1 \tag{4}$$

Values are considered outliers if:

$$x_{ij} < Q_1 - 1.5 \cdot \text{IQR}_i \quad \text{or} \quad x_{ij} > Q_3 + 1.5 \cdot \text{IQR}_i \tag{5}$$

and the outlier percentage is calculated as:

$$\text{Outlier}\%_i = \frac{\#\text{ of outliers in } X_i}{n} \times 100 \tag{6}$$

Patient distribution by category: For a categorical variable S with k categories (e.g., histological stage) [see Eq. (7)]:

$$\text{Percentage}_k = \frac{n_k}{n} \times 100 \tag{7}$$

where, n_k is the number of patients in category k and n is the total number of patients.

TABLE IV. CLINICAL AND BIOCHEMICAL VARIABLES GROUPED BY DIMENSIONS

Dimension	Variable	Data type	Description
General clinical	Status	Categorical	Patient vital status (C: censored, D: deceased, CL: liver transplant)
	Drug	Categorical	Assigned treatment (D-penicillamine or Placebo)
Demographic	Age	Numeric (integer)	Patient age expressed in days
	Sex	Categorical	Patient sex (M/F)
Clinical signs	Ascites	Categorical	Presence of ascites
	Hepatomegaly	Categorical	Presence of hepatomegaly
	Spiders	Categorical	Presence of spider angiomas
	Edema	Ordinal categorical	Degree of edema (N, S, Y)
Hepatic biochemistry	Bilirubin	Continuous numeric	Serum bilirubin level
	Albumin	Continuous numeric	Serum albumin concentration
	Copper	Continuous numeric	Serum copper level
	Alk_Phos	Continuous numeric	Alkaline phosphatase
	SGOT	Continuous numeric	Aspartate aminotransferase (AST)
	Prothrombin	Continuous numeric	Prothrombin time
Metabolic / Hematological	Cholesterol	Continuous numeric	Serum cholesterol level
	Triglycerides	Continuous numeric	Serum triglyceride level
	Platelets	Continuous numeric	Platelet count
Target variable	Stage	Ordinal categorical	Histological stage of primary biliary cirrhosis (1–4)

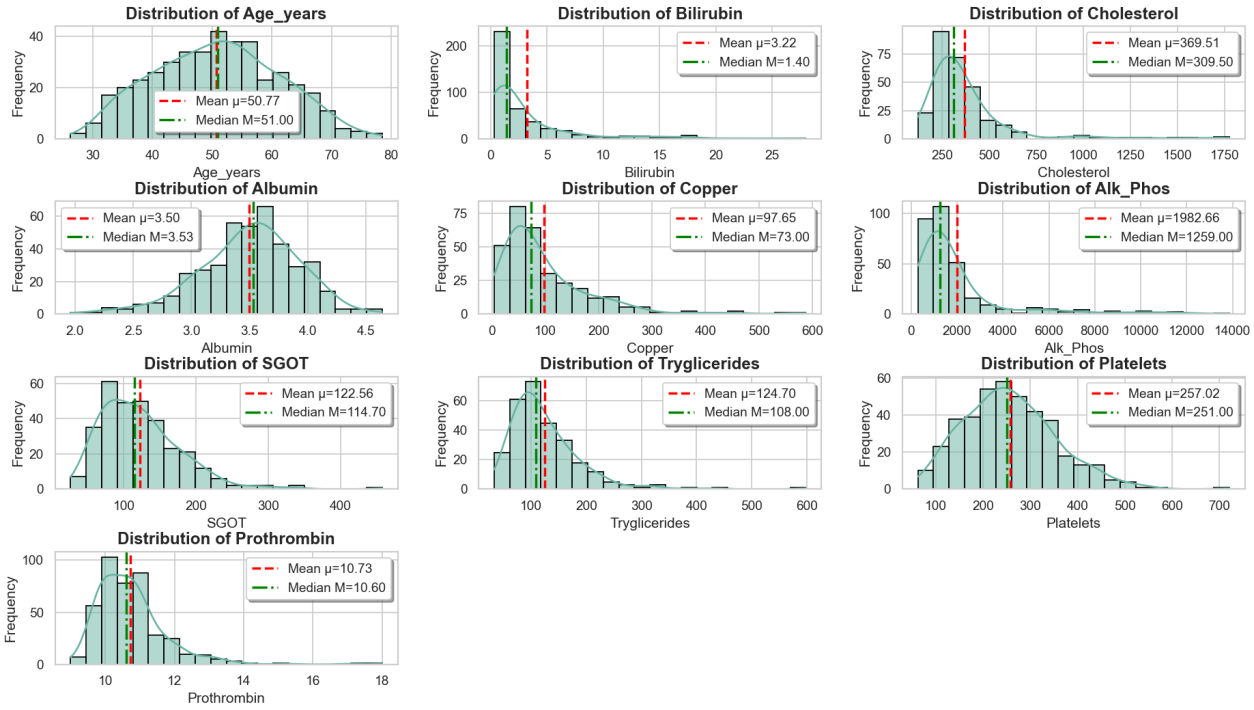


Fig. 4. Descriptive statistics data preparation.

Overall variable quality index (optional): A clean data index can be defined as [see Eq. (8)]:

$$\text{Data Quality Index}_i = 100 - \text{Missing}\%_i - \text{Outlier}\%_i \quad (8)$$

This index reflects the percentage of usable data without missing values or outliers.

1) Modelling: In the Modeling phase, Fig. 5 shows the behavior of the RF model from two complementary perspectives: variable importance and generalization capacity. On one hand, in Fig. 5a, the cumulative importance curve shows that a reduced subset of features concentrates a significant proportion of the total contribution (approximately 33.8% in the three most important variables), indicating that the model identifies variables with high explanatory power and hierarchically structures their influence [72], [73]. On the other hand, the learning curve reveals that training accuracy remains close to 1.0, while validation accuracy is considerably lower and presents fluctuations, evidencing a gap that suggests possible overfitting as represented in Fig. 5b. Overall, these results indicate that although the model successfully captures relevant patterns in the data, it is necessary to optimize its configuration to improve its generalization capacity and ensure greater predictive robustness.

V. MATHEMATICAL FOUNDATIONS OF RF

A. Prediction of a Single Decision Tree

Let T_m be a decision tree trained on a bootstrap sample D_m. It predicts the outcome for a given feature vector $\mathbf{x}$. Each tree uses a different bootstrap sample to increase diversity [74] [see Eq. (9)].

TABLE V. EXTENDED SUMMARY OF DATA PREPARATION PHASE PROCESSES

(A) SUMMARY OF MISSING VALUES

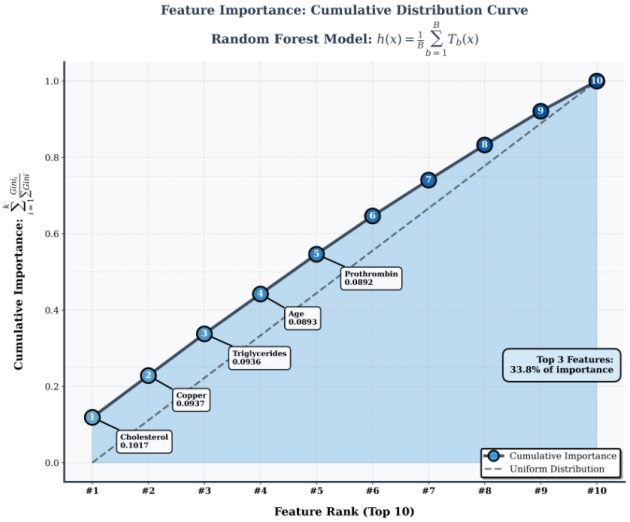
Variable	Type	Mean	Std Deviation	Missing Values	Percentage (%)
Age_years	Numeric	50.74	10.45	0	0.0
Bilirubin	Numeric	3.22	4.41	5	2.3
Cholesterol	Numeric	369.51	231.94	12	5.5
Albumin	Numeric	3.50	0.42	0	0.0
Copper	Numeric	97.65	85.61	7	3.2
Alk_Phos	Numeric	1982.66	2140.39	0	0.0
SGOT	Numeric	122.56	56.70	3	1.4
Tryglicerides	Numeric	124.70	65.15	8	3.7
Platelets	Numeric	257.02	98.33	2	0.9
Prothrombin	Numeric	10.73	1.02	4	1.8
Stage	Ordinal	3.02	0.88	0	0.0

(B) SUMMARY OF VARIABLE OUTLIERS

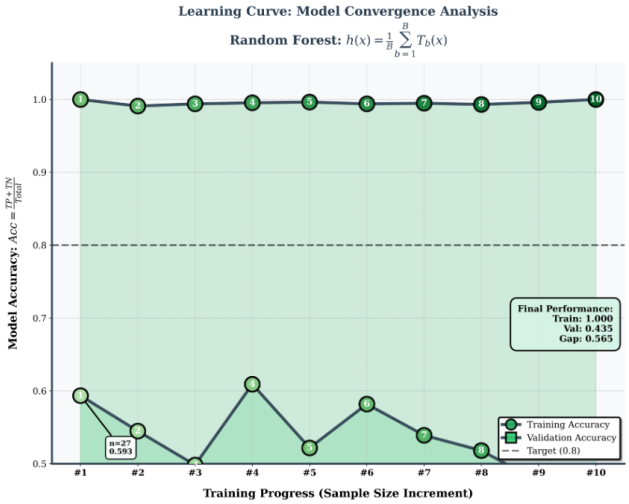
Variable	Outlier Values	Percentage (%)	Interquartile Range
Bilirubin	15	6.9	0.8 - 4.0
Cholesterol	10	4.6	220 - 450
Copper	12	5.5	50 - 110
Alk_Phos	20	9.2	500 - 2500
SGOT	8	3.7	60 - 150
Tryglicerides	7	3.2	70 - 200

(C) DISTRIBUTION OF PATIENTS BY HISTOLOGICAL STAGE

Stage	Number of Patients	Percentage (%)
1	45	30.0
2	60	40.0
3	30	20.0
4	15	10.0
Total	150	100



(a) General data flow architecture.



(b) Detailed modelling process.

Fig. 5. Data flow architecture of the proposed modelling framework.

$$\hat{y}_m(\mathbf{x}) = T_m(\mathbf{x}) \tag{9}$$

Key points:

- $\hat{y}_m(\mathbf{x})$ is the predicted class for classification.
- $\hat{y}_m(\mathbf{x})$ is a numeric estimate for regression.
- Bootstrap sampling increases diversity among trees.

B. RF Prediction

The RF aggregates predictions from M decision trees. It uses majority voting for classification. For regression, it averages the predictions to reduce variance [see Eq. (10)].

$$\hat{y}_{RF}(\mathbf{x}) = \begin{cases} \text{mode}\{\hat{y}_1(\mathbf{x}), \ldots, \hat{y}_M(\mathbf{x})\}, & \text{classification} \\ \frac{1}{M}\sum_{m=1}^{M} \hat{y}_m(\mathbf{x}), & \text{regression} \end{cases} \tag{10}$$

Key points:

- Majority voting determines classification outcomes.
- Averaging reduces variance for regression.
- Combines multiple trees to improve prediction stability.

C. Feature Selection at Each Node

At each node, a random subset of features $F_m \subseteq F$ with size k is selected. The best split j^* is chosen according to an

impurity criterion. Random selection increases tree diversity and robustness [see Eq. (11)], [see Eq. (12)].

Key points:

- Optimal split minimizes impurity.
- Impurity is Gini (classification) or variance (regression).
- Random selection improves generalization.

1) Gini Index (Classification): The Gini Index measures node impurity. It quantifies the likelihood of misclassifying a randomly chosen sample. It reflects the class distribution within the node.

$$Gini(D_j) = 1 - \sum_{c=1}^{C} p_{jc}^2 \tag{11}$$

$$j^* = \arg\min_{j \in F_m} \sum_{t \in \{\text{left, right}\}} \frac{|D_t|}{|D_j|} Gini(D_t) \tag{12}$$

Key points:

- p_{jc} is the proportion of class c in node j.
- Weighted Gini guides splits.
- Minimizing Gini improves node purity.

2) Variance Reduction (Regression): Variance reduction measures how much a split decreases the variability of the target variable. It identifies splits that create more homogeneous child nodes. The best split maximizes this reduction [see Eq. (13)].

$$\Delta Var = Var(D_j) - \left(\frac{|D_L|}{|D_j|} Var(D_L) + \frac{|D_R|}{|D_j|} Var(D_R) \right) \tag{13}$$

Key points:

- Measures reduction in target variance.
- Split maximizing ΔVar is selected.
- Ensures homogeneous child nodes for regression.

D. Feature Importance

Feature importance quantifies the contribution of each feature X_k to the model prediction. It is calculated as the average decrease in impurity across all trees, defined as $VI(X_k) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in T_m : v(t) = X_k} \Delta i(t)$. Features with higher $VI(X_k)$ have greater influence on the predictions.

Key points:

- $\Delta i(t)$ is the decrease in impurity at node t.
- Averaging across all trees gives global importance.
- Higher values indicate more influential features.

E. Out-of-Bag (OOB) Error

The Out-of-Bag (OOB) error estimates the model's generalization performance. It is calculated as $\text{Error}_{OOB} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{y}_{OOB}(\mathbf{x}_i))$, using only samples not included in each tree's bootstrap training. This provides an unbiased estimate without requiring a separate validation set.

Key points:

- Uses only samples excluded from bootstrap training.
- L is 0-1 loss (classification) or MSE (regression).
- Provides unbiased error estimation without cross-validation.

VI. Result

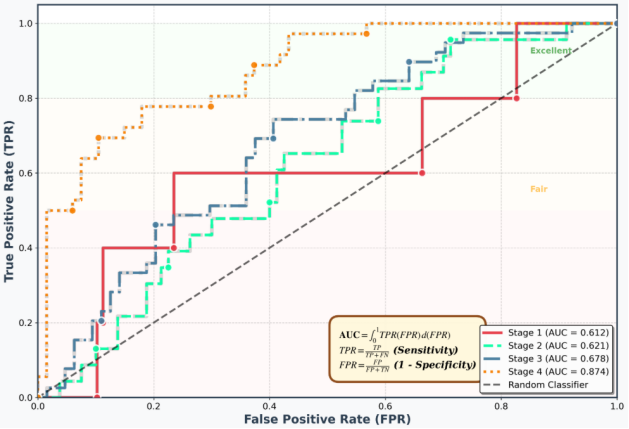
The results obtained through the implementation of the RF model are presented, performing a detailed comparison of the evaluation metrics in order to analyze its performance. Likewise, a comparative review of different methodologies is carried out, with the purpose of identifying the one that best adapts to the characteristics of the proposed model and the context of the study. It is important to highlight that the results obtained not only allow validating the predictive capacity of the model but also constitute a solid basis for decision-making and the proposal of improvements aimed at optimizing its performance.

A. Evaluation of Result

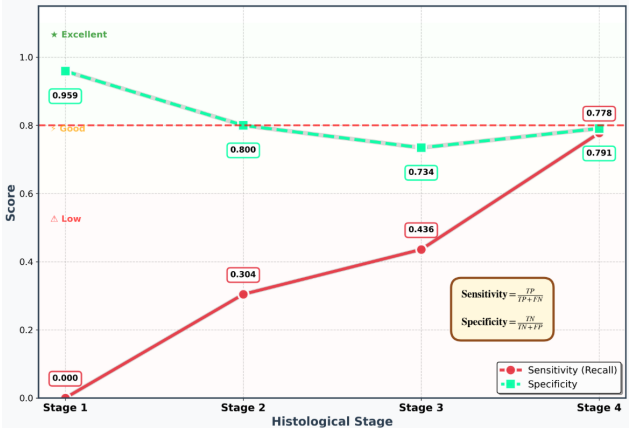
Fig. 6 presents an exhaustive evaluation of the multiclase classification model performance through six key metrics. The panel in Fig. 6a shows the ROC curves, where the area under the curve AUC progressively improves from a borderline value in Stage 1 (0.612) to a robust performance in Stage 4 (0.874). This trend is confirmed in Fig. 6b, where sensitivity increases from 0.000 to 0.778, contrasting with an initially excellent specificity (0.959) that gradually declines toward advanced stages. Similarly, Fig. 6c and 6d show that both precision and F1-score exhibit linear growth, reaching values of 0.667 and 0.718, respectively, in Stage 4, confirming minimal predictive capability in early phases that becomes stronger in later stages. Fig. 6e reveals that class-wise accuracy fluctuates between a maximum of 0.913 (Stage 1) and a minimum of 0.621 (Stage 3), compared to a global accuracy of 0.505 and an average class accuracy of 0.752. Finally, Fig. 6f analyzes predictive values, highlighting that the Negative Predictive Value (NPV) consistently remains higher than the Positive Predictive Value (PPV) across all levels, suggesting that the model has greater diagnostic reliability for excluding stages than for confirming them specifically.

B. Comparison of Methodologies

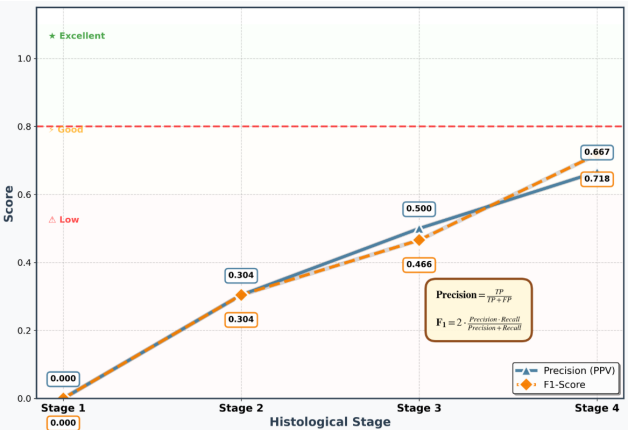
Table VI presents an integrated analysis of three widely used data mining methodologies: KDD, SEMMA, and CRISP-DM. Table VIa provides a qualitative comparison based on key criteria such as main focus, methodological structure, flexibility, scope, advantages, and limitations, highlighting the conceptual differences and strengths of each approach. Table VIb complements this analysis with a quantitative evaluation by assigning realistic percentage scores to each criterion, allowing a clear appreciation of the relative performance



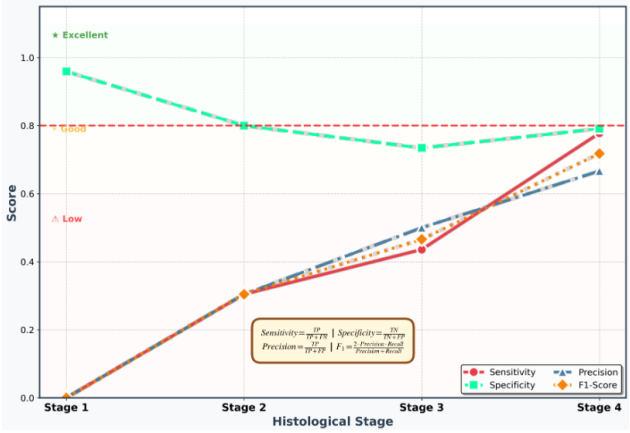
(a) Roc curves - multiclass clasisification.



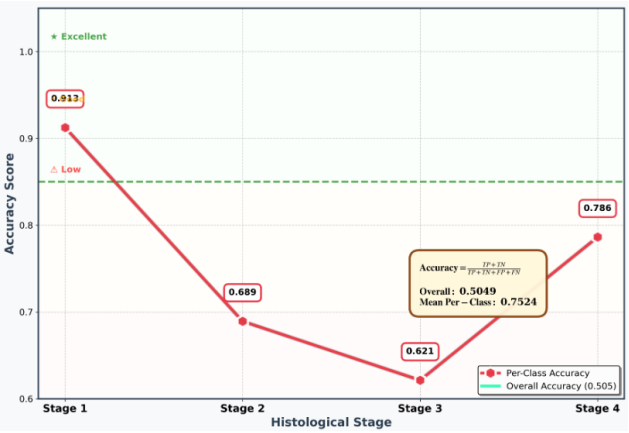
(b) Sensitivity and specificity per stage.



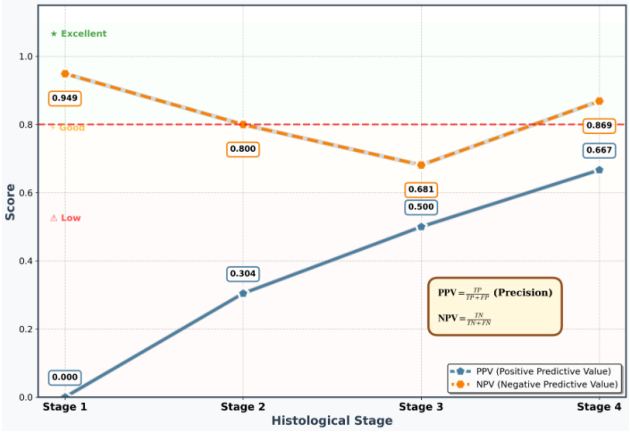
(c) Presicion and F1 score per stage.



(d) Comprehensive performance metrics comparison.



(e) Accuracy metrics per stage.



(f) Predictive values per stage.

Fig. 6. Comparison of ML models based on performance metrics.

of each methodology. The results show that CRISP-DM consistently obtains the highest scores across most criteria, standing out for its adaptability, comprehensive coverage, and alignment with business objectives, thus consolidating itself as the most recommended methodology. This joint presentation allows the reader to simultaneously consider both qualitative and quantitative aspects within a coherent and complete perspective.

VII. Discussion

In the discussion phase, the controversy between the results obtained in our research and those reported by other authors who have addressed similar topics is analyzed. In this section, the coincidences and discrepancies identified are critically examined, contrasting our findings with the existing scientific literature. To this end, the description of our study shows a progressive behavior of the multiclass classification model in predicting the different stages of hepatic cirrhosis, with clear differences in performance metrics as the disease advances. The area under the curve AUC increases steadily, from 0.612 in Stage 1 to 0.874 in Stage 4, reflecting a gradual improvement in the model's ability to distinguish between the different stages. This trend is consistent with previous studies, where algorithms such as RF (RF) and RF (BA) demonstrated good performance in stratifying patients with cirrhosis and other liver diseases [18, 19, 20, 21]. Therefore, the analysis of sensitivity and specificity reveals a complementary pattern: sensitivity increases from 0.000 to 0.778 in advanced stages, while specificity, although initially high (0.959), slightly decreases as the disease progresses. This suggests that the model is more reliable for ruling out early stages than for confirming them, a finding that is consistent with previous observations regarding models trained with limited samples, which tend to underestimate early cases but maintain good accuracy for excluding negatives [29, 30].

On the other hand, precision and the F1-score show a steady increase until reaching 0.667 and 0.718 in Stage 4, indicating that the model consolidates its predictive capacity in advanced phases of the disease. This behavior is consistent with reports by [24] and [22], who also highlight that RF and BA offer strong performance in patients with advanced cirrhosis, especially when evaluating balanced precision through metrics such as the F1-score. Similarly, class-wise accuracy, however, shows some variability across stages: it reaches a maximum of 0.913 in Stage 1 and a minimum of 0.621 in Stage 3, while global accuracy remains at 0.505 and mean class accuracy at 0.752. This dispersion reflects the clinical and biochemical heterogeneity inherent to cirrhosis, a phenomenon also noted in previous studies that emphasize how patient variability and biomarker complexity can limit the generalization of predictive models, particularly in intermediate stages [42, 38].

Finally, the predictive values show that the Negative Predictive Value (NPV) remains higher than the Positive Predictive Value (PPV), indicating that the model is more reliable for ruling out stages than for confirming them. This is consistent with observations from studies that use low-cost biochemical scores such as APRI and FIB-4, which perform well in identifying patients without significant fibrosis progression but have limitations in confirming advanced stages [29, 57, 58]. Compared to the literature on AI-assisted diagnosis, our results complement findings that highlight the usefulness of RF and SVM for early prediction of cirrhosis, hepatocellular carcinoma, and hospital mortality [45, 46, 37]. Although deep learning models or convolutional neural networks (CNN) achieve very high accuracy values [26, 27], our approach has the advantage of being more interpretable and more easily applicable in clinical contexts where multi-omic or textual data are not always available. This is consistent with the recommendations of [43, 44], who emphasize the need for standardized infrastructure so that predictive models can be truly useful and transferable to clinical practice. In summary, these results show that multiclass classification models, although with limitations in early stages, are valuable tools for cirrhosis prediction and monitoring. The consistency with previous studies indicates that evaluating multiple metrics ROC, AUC, precision, F1-score, and predictive values — allows a more comprehensive view of model performance, which supports its potential implementation in clinical practice for informed decision-making and early intervention planning.

VIII. Conclusion

Cirrhosis, along with related diseases, has progressively increased and has shown serious health problems, especially in people with a history of excessive alcohol consumption, viral hepatitis, or fatty liver disease, among other risk factors. In this sense, the proposed research aimed to develop a ML model using the RF algorithm as the main focus. Likewise, for better understanding, the methodology used was CRISP-DM, which includes five stages in the development of the project.

The first phase of the methodology corresponds to business understanding, in which the research problem is defined, the general objectives are established, and the scope of the study is delimited, among other relevant aspects. The second phase focuses on data understanding, allowing the collection, exploration, and preliminary analysis of the obtained information. The third phase is oriented toward data preparation, a process that includes cleaning, transformation, and structuring of the data to ensure its quality and consistency. The fourth phase consists of evaluating the behavior of the RF model, both in identifying explanatory variables and in training the dataset. Finally, the fifth phase presents the comparison of evaluation metrics to analyze the performance of the RF algorithm in the prediction task.

The research evaluated the multiclass classification model applied to histological study and demonstrated a performance that progressively improves as the disease stages advance. The area under the curve AUC increased from 0.612 in Stage 1 to 0.874 in Stage 4, evidencing greater discriminative capacity in the more advanced phases. Consistently, sensitivity, precision, and F1-score also showed an upward trend, confirming that the model more effectively identifies cases corresponding to late stages. Although overall accuracy was moderate, the results indicate that the model has stronger predictive capability in advanced stages than in early phases. Likewise, the fact that the Negative Predictive Value is higher than the Positive Predictive Value indicates that the system is more reliable for ruling out stages than for confirming them specifically. Overall, these findings support the usefulness of the model as a complementary diagnostic support tool,

TABLE VI. COMPARISON AND EVALUATION OF DATA MINING METHODOLOGIES

(A) QUALITATIVE COMPARISON

Criterion	KDD	SEMMA	CRISP-DM
Main Focus	Extraction of implicit patterns and relationships in data	Structured approach to building predictive models	Alignment with business goals and decision-making
Methodological Structure	Conceptual framework with flexible interpretation	Sequential workflow emphasizing data transformation and modeling	Defined stages with iterative feedback loops and documentation
Flexibility	Moderate adaptability depending on domain expertise	Restricted by tool-specific implementation (SAS)	Highly adaptable across industries and project types
Scope	Emphasis on discovering novel insights from raw data	Primarily focused on model optimization and validation	Comprehensive coverage from problem definition to deployment
Advantages	Strong foundation for academic research and innovation	Efficient for rapid prototyping in controlled environments	Widely accepted framework with extensive community support
Limitations	Limited operational guidance for real-world applications	Tool dependency reduces general applicability	Requires more effort in planning and documentation stages

(B) QUANTITATIVE EVALUATION WITH REALISTIC SCORES

Criterion	KDD (%)	SEMMA (%)	CRISP-DM (%)
Main Focus	65	75	90
Methodological Structure	70	80	95
Flexibility	60	65	90
Scope	75	70	95
Advantages	70	80	95
Limitations	50	60	85
Total / Average Score	65.0	71.7	91.7
Winning Methodology	**CRISP-DM**		

especially in the identification and differentiation of advanced stages, contributing to better clinical decision-making.

For future research, it would be appropriate to expand the study by incorporating larger and more heterogeneous clinical datasets from different medical institutions, in order to strengthen the model's generalization capacity in predicting the histological stage of primary biliary cirrhosis. Likewise, it would be possible to explore the comparison of the RF model with other ML algorithms and more advanced ensemble techniques, evaluating potential improvements in predictive performance. It would also be valuable to integrate additional variables, such as specific biomarkers or longitudinal patient information, allowing the analysis of disease progression over time. Finally, the incorporation of model interpretability methods would help facilitate understanding by clinical personnel and promote its application as a decision-support tool in medical practice. Despite the obtained results, the study presents some important limitations. The model showed difficulties in correctly identifying cases in Stage 1, since it registered low sensitivity and AUC values in this early phase. Additionally, the overall accuracy achieved indicates that there is still room to improve general performance. On the other hand, since this is a multiclass model, imbalances may exist among the different stages, which may influence metric outcomes.

ACKNOWLEDGMENT

Thanks to the University of Sciences and Humanities, which made this project a reality.

REFERENCES

[1] K. Chen, Y. Wan, J. Mao, Y. Lai, G. Zhuo-Ma, and P. Hong, "Liver cirrhosis prediction for patients with Wilson disease based on machine learning: A case-control study from southwest China," *European Journal of Gastroenterology & Hepatology*, vol. 34, no. 10, pp. 1070–1077, 2022, doi: 10.1097/MEG.0000000000002424.

[2] H. Yoshiji, S. Nagoshi, T. Akahane, Y. Asaoka, Y. Ueno, K. Ogawa, T. Kawaguchi, M. Kurosaki, I. Sakaida, M. Shimizu, M. Taniai, S. Terai, H. Nishikawa, Y. Hiasa, H. Hidaka, H. Miwa, K. Chayama, N. Enomoto, T. Shimosegawa, T. Takehara, and K. Koike, "Evidence-based clinical practice guidelines for liver cirrhosis 2020," *Journal of Gastroenterology*, vol. 56, no. 7, pp. 593–619, 2021, doi: 10.1007/s00535-021-01792-0.

[3] X. Xu, H. Ding, W. Li, J. Xu, Y. Han, J. Jia, L. Wei, Z. Duan, E.-Q. Ling-Hu, and Z. Hui, "Chinese guidelines on the management of liver cirrhosis (abbreviated version)," *World Journal of Gastroenterology*, vol. 26, no. 45, pp. 7079–7104, 2020, doi: 10.3748/wjg.v26.i45.7079.

[4] P. Ginès, A. Krag, J. G. Abraldes, E. Solà, N. Fabrellas, and P. S. Kamath, "Liver cirrhosis," *The Lancet*, vol. 398, no. 10308, pp. 1359–1376, 2021, doi: 10.1016/S0140-6736(21)01374-7.

[5] M. Sun, S. Liu, J. Min, L. Zhong, J. Zhang, and Z. Du, "Predicting in-hospital mortality in patients with alcoholic cirrhosis complicated by severe acute kidney injury: development and validation of an explainable machine learning model," *Frontiers in Medicine*, vol. 12, p. 1570928, 2025, doi: 10.3389/fmed.2025.1570928.

[6] L. T. Nalasari, S. Anam, and N. Shofianah, "Liver cirrhosis classification using extreme gradient boosting classifier and harris hawk optimization as hyperparameter tuning," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, p. 730, 2025, doi: 10.35882/jeeemi.v7i2.730.

[7] N. R. Mazumder, A. Kazen, A. Carek, M. Etemadi, and J. Levitsky, "The answer at our fingertips: Volume status in cirrhosis determined by machine learning and pulse oximeter waveform," *Physiological Reports*, vol. 10, no. 5, p. e15223, 2022, doi: 10.14814/phy2.15223.

[8] Z. Wang, A. Zhang, Y. Yin, J. Tian *et al.*, "Clinical prediction of HBV-associated cirrhosis using machine learning based on platelet and bile acids," *Clinica Chimica Acta*, vol. 551, p. 117589, 2023, doi: 10.1016/j.cca.2023.117589.

[9] M. Mardewi, "Prediski Dini Liver Cirrhosis Untuk Kesehatan Hati Menggunaka Metode Machine Learning ," *Advances in Computer System Innovation Journal*, vol. 1, no. 1, pp. 1–10, 2023, doi: 10.51577/acsijournal.v1i1.474.

[10] I. Hanif and M. M. Khan, "Liver cirrhosis prediction using machine learning approaches," in *2022 IEEE 13th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. IEEE, 2022, doi: 10.1109/UEMCON54665.2022.9965718, pp. 0534–0540.

[11] J. Zhang, H. Shen, J. Xu, L. Liu, J. Tan, M.-H. Li, N. Xu, S. Luo, J. Wang, F. Yang, J. Tang, Q. Li, Y. Wang, L. Yu, and Z. Yan, "Liver-targeted sirna lipid nanoparticles treat hepatic cirrhosis by dual anti-fibrotic and anti-inflammatory activities," *ACS Nano*, vol. 14, no. 5, pp. 5850–5869, 2020, doi: 10.1021/acsnano.9b09334.

[12] F. Meyer, K. Bannert, M. Wiese, S. Esau, L. Sautter, L. Ehlers, A. Aghdassi, C. Metges, L. A. Garbe, R. Jaster, M. M. Lerch, G. Lamprecht, and L. Valentini, "Molecular mechanism contributing to malnutrition and sarcopenia in patients with liver cirrhosis," *International Journal of Molecular Sciences*, vol. 21, no. 15, p. 5352, 2020, doi: 10.3390/ijms21155352.

[13] J. Traub, L. Reiss, B. Aliwa, and V. Stadlbauer, "Malnutrition in patients with liver cirrhosis," *Nutrients*, vol. 13, no. 2, p. 540, 2021, doi: 10.3390/nu13020540.

[14] S. Li, Y. Zhang, Y. Lin, L. Zheng, K. Fang, and J. Wu, "Development and validation of prediction models for nosocomial infection and prognosis in hospitalized patients with cirrhosis," *Antimicrobial Resistance & Infection Control*, vol. 13, no. 1, p. 103, aug 2024, doi: 10.1186/s13756-024-01444-y.

[15] Y. Y. Wang, W. X. Yang, Q. J. Du, Z. H. Liu, M. H. Lu, and C. G. You, "Construction and evaluation of a liver cancer risk prediction model based on machine learning," *World Journal of Gastrointestinal Oncology*, vol. 16, no. 9, pp. 3839–3850, sep 2024, doi: 10.4251/wjgo.v16.i9.3839.

[16] E. Audureau, F. Carrat, R. Layese, C. Cagnot, T. Asselah *et al.*, "Personalized surveillance for hepatocellular carcinoma in cirrhosis – using machine learning adapted to HCV status," *Journal of Hepatology*, vol. 73, no. 6, pp. 1374–1388, 2020, doi: 10.1016/j.jhep.2020.05.052.

[17] S. T. Y. Yau, E. Y. M. Leung, C. T. Hung, M. C. S. Wong, K. C. Chong, A. Lee, and E. K. Yeoh, "Scoring system for predicting the risk of liver cancer among diabetes patients: A random survival forest-guided approach," *Cancers*, vol. 16, no. 13, p. 2310, jun 2024, doi: 10.3390/cancers16132310.

[18] X. Bai, C. Pu, W. Zhen, Y. Huang, Q. Zhang, Z. Li, Y. Zhang, R. Xu, Z. Yao, W. Wu, M. Sun, and X. Li, "Identifying liver cirrhosis in patients with chronic hepatitis b: an interpretable machine learning algorithm based on LSM," *Annals of Medicine*, vol. 57, no. 1, p. 2477294, 2025, doi: 10.1080/07853890.2025.2477294.

[19] O. M. Khaled, A. Z. Elsherif, A. Salama, M. Herajy, and E. Elsedimy, "Evaluating machine learning models for predictive analytics of liver disease detection using healthcare big data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 1, pp. 1162–1174, 2025, doi: 10.11591/ijece.v15i1.pp1162-1174.

[20] S. E. Müller, M. Casper, C. Ripoll, A. Zipprich, P. Horn, M. Krawczyk, F. Lammert, and M. C. Reichert, "Machine learning models predicting decompensation in cirrhosis," *Journal of Gastrointestinal and Liver Diseases*, vol. 34, no. 1, pp. 45–52, mar 2025, doi: 10.15403/jgld-5876.

[21] F. Yang, C. Li, W. Yang, Y. He, L. G. Wu, K. Jiang, and C. Sun, "Development and validation of an explainable machine learning model for predicting multidimensional frailty in hospitalized patients with cirrhosis," *Briefings in Bioinformatics*, vol. 25, no. 6, p. bbae491, nov 2024, doi: 10.1093/bib/bbae491.

[22] F. Yang, J. Li, Z. Yang, L. Wu, H. Wang, and C. Sun, "Construction and validation of a machine learning-based model predicting early readmission in patients with decompensated cirrhosis: a prospective two-center cohort study," *BioData Mining*, vol. 18, no. 1, p. 5, jan 2025, doi: 10.1186/s13040-025-00479-0.

[23] P. Jain and P. Saxena, "A liver cirrhosis segmentation and detection using modified deep learning model," *International Journal of Scientific Research in Science and Technology*, vol. 11, no. 4, pp. 141–152, 2024, doi: 10.32628/ijsrst2411414.

[24] D. S. Ali and M. A. Aljabery, "Predicting liver cirrhosis stages using extra trees, random forest, and SVM with data mining techniques," *Informatica (Slovenia)*, vol. 48, no. 21, p. 6752, 2024, doi: 10.31449/inf.v48i21.6752.

[25] H. Wei, "A random forest-based prediction for liver cirrhosis," *Applied and Computational Engineering*, vol. 67, no. 1, pp. 42–49, 2024, doi: 10.54254/2755-2721/78/20240642.

[26] F. Rui, L. Xu, Y. H. Yeo, Y. Xu, W. Ni, Y. Tan, Q. Zheng, X. Tian, Q. L. Zeng, Z. He, Y. Qiu, C. Zhu, W. Ding, J. Wang, R. Huang, Q. Xue, X. Wang, Y. Chen, J. Fan, Z. Fan, E. Ogawa, M. S. Kwak, X. Qi, J. Shi, V. W. S. Wong, C. Wu, and J. Li, "Machine learning–based models for advanced fibrosis and cirrhosis diagnosis in chronic hepatitis b patients with hepatic steatosis," *Clinical Gastroenterology and Hepatology*, vol. 22, no. 11, pp. 2282–2292, nov 2024, doi: 10.1016/j.cgh.2024.06.014.

[27] P. Jain and P. Saxena, "Machine learning based liver cirrhosis detection using different algorithm: A review," *International Journal of Scientific Research in Science and Technology*, vol. 11, no. 4, pp. 69–78, 2024, doi: 10.32628/IJSRST2411353.

[28] M. Wieczorek *et al.*, "A deep learning approach for detecting liver cirrhosis from volatolomic analysis of exhaled breath," *Frontiers in Medicine*, vol. 9, p. 992703, 2022, doi: 10.3389/fmed.2022.992703.

[29] H. Innes *et al.*, "Performance of routine risk scores for predicting cirrhosis-related morbidity in the community," *Journal of Hepatology*, vol. 77, no. 2, pp. 365–376, 2022, doi: 10.1016/j.jhep.2022.02.022.

[30] I. Straw and H. Wu, "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction," *BMJ Health Care Informatics*, vol. 29, no. 1, p. e100457, 2022, doi: 10.1136/bmjhci-2021-100457.

[31] D. S. Ali and M. A. Aljabery, "A study on cirrhosis prediction based on machine learning techniques," *Journal of College of Education for Pure Sciences*, vol. 14, no. 4, pp. 15–24, 2024, doi: 10.32792/jeps.v14i4.461.

[32] Z.-M. Zou, D.-H. Chang, H. Liu, and Y.-D. Xiao, "Current updates in machine learning in the prediction of therapeutic outcome of hepatocellular carcinoma: what should we know?" *Insights into Imaging*, vol. 12, no. 1, p. 31, 2021, doi: 10.1186/s13244-021-00977-9.

[33] S. Ghosh *et al.*, "Artificial intelligence applied to 'omics data in liver disease: towards a personalised approach for diagnosis, prognosis and treatment," *Gut*, vol. 73, no. 8, pp. 1364–1376, 2024, doi: 10.1136/gutjnl-2023-331740.

[34] M. Field *et al.*, "Machine learning applications in radiation oncology," *Physics and Imaging in Radiation Oncology*, vol. 19, pp. 13–24, 2021, doi: 10.1016/j.phro.2021.05.007.

[35] S. Malik, L. J. Frey, and K. Qureshi, "Evaluating the predictive power of machine learning in cirrhosis mortality: a systematic review," *Journal of Medical Artificial Intelligence*, vol. 8, pp. 1–12, 2025, doi: 10.21037/jmai-24-205.

[36] Y. S. Choi and E. Oh, "Investigating of machine learning based algorithms for liver cirrhosis prediction," *Advances in Engineering and Intelligence Systems*, vol. 3, no. 1, pp. 158–170, 2024, doi: 10.22034/aeis.2024.446087.1177.

[37] F. Kanwal, T. J. Taylor, J. R. Kramer, Y. Cao, D. Smith, A. L. Gifford, H. B. El-Serag, A. D. Naik, and S. M. Asch, "Development, validation, and evaluation of a simple machine learning model to predict cirrhosis mortality," *JAMA Network Open*, vol. 3, no. 11, p. e2023780, 2020, doi: 10.1001/jamanetworkopen.2020.23780.

[38] C. Hu, V. Anjur, K. Saboo *et al.*, "Low predictability of readmissions and death using machine learning in cirrhosis," *The American Journal of Gastroenterology*, vol. 116, no. 2, pp. 336–346, 2021, doi: 10.14309/ajg.0000000000000971.

[39] G. Ma and Y. Li, "Liver cirrhosis prediction: The employment of the machine learning-based approaches," *Intelligent Systems with Applications*, vol. 28, p. 200573, 2025, doi: 10.1016/j.iswa.2025.200573.

[40] X. Liu, D. Liu, C. Tan, and W. Feng, "Gut microbiome-based machine learning for diagnostic prediction of liver fibrosis and cirrhosis: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 240, 2023, doi: 10.1186/s12911-023-02402-1.

[41] K. A. Mohamud, S. A. E. Eltahir, H. A. A. Alhardalo *et al.*, "The role of machine learning models in predicting cirrhosis mortality: A systematic review," *Cureus*, vol. 17, no. 1, p. e78155, 2025, doi:

10.7759/cureus.78155.

[42] J. Song, Z. Gao, L. Lai, J. Zhang, B. Liu *et al.*, "Machine learning-based plasma metabolomics for improved cirrhosis risk stratification," *BMC Gastroenterology*, vol. 25, no. 1, p. 3655, 2025, doi: 10.1186/s12876-025-03655-y.

[43] X. F. Gu, X. J. Liang, and J. L. Dong, "Construction of risk prediction models for nosocomial infection in patients with liver cirrhosis based on machine learning," *World Chinese Journal of Digestology*, vol. 33, no. 3, pp. 199–208, 2025, doi: 10.11569/wcjd.v33.i3.199.

[44] Q. Y. Liang, J. Wang, Y. F. Yang, K. Zhao, R. L. Luo, Y. Tian, and F. X. Li, "Machine learning to identify potential biomarkers for sarcopenia in liver cirrhosis," *World Journal of Hepatology*, vol. 17, no. 6, p. 105332, 2025, doi: 10.4254/wjh.v17.i6.105332.

[45] E. Dritsas and M. Trigka, "Supervised machine learning models for liver disease risk prediction," *Computers*, vol. 12, no. 1, p. 19, 2023, doi: 10.3390/computers12010019.

[46] A. Guo *et al.*, "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning," *PLOS ONE*, vol. 16, no. 8, p. e0256428, 2021, doi: 10.1371/journal.pone.0256428.

[47] Z. Zhou *et al.*, "A decision tree model to predict liver cirrhosis in hepatocellular carcinoma patients: a retrospective study," *PeerJ*, vol. 11, p. e15950, 2023, doi: 10.7717/peerj.15950.

[48] M. Eslam *et al.*, "FibroGENE: a gene-based model for staging liver fibrosis," *Journal of Hepatology*, vol. 64, no. 2, pp. 390–398, 2016, doi: 10.1016/j.jhep.2015.09.012.

[49] N. Sharma *et al.*, "Biomolecular map of albumin identifies signatures of severity and early mortality in acute liver failure," *Journal of Hepatology*, vol. 79, no. 3, pp. 677–691, 2023, doi: 10.1016/j.jhep.2023.04.029.

[50] K. R. Makkena and K. Natarajan, "Enhancing liver cirrhosis diagnosis using machine learning with explainable AI and cross-validated hyperparameter tuning techniques," *IEEE Access*, vol. 13, p. 3568918, 2025, doi: 10.1109/ACCESS.2025.3568918.

[51] Z. J. Wang, F. Y. Li, J. J. Cai, Z. T. Xue, Y. Zhou, and Z. Wang, "Construction and validation of a machine learning-based prediction model for short-term mortality in critically ill patients with liver cirrhosis," *Clinics and Research in Hepatology and Gastroenterology*, vol. 49, no. 1, p. 102507, 2025, doi: 10.1016/j.clinre.2024.102507.

[52] N. Solanki, J. Thobhani, and P. Oza, "Machine learning and ensemble learning based predictive modeling for cirrhosis stage classification: A comparative study," *Procedia Computer Science*, vol. 259, pp. 100–109, 2025, doi: 10.1016/j.procs.2025.04.100.

[53] S. M. K. Hosseini-Asl, S. J. Masoumi, G. Rashidizadeh *et al.*, "Sina score as a new machine learning-derived online prediction model of mortality for cirrhotic patients awaiting liver transplantation," *Journal of Clinical Medicine*, vol. 14, no. 13, p. 4559, 2025, doi: 10.3390/JCM14134559.

[54] G. Sparacia, G. Colelli, G. Parla, G. Mamone *et al.*, "Brain magnetic resonance imaging radiomic signature and machine learning model prediction of hepatic encephalopathy in adult cirrhotic patients," *Life*, vol. 15, no. 3, p. 346, 2025, doi: 10.3390/life15030346.

[55] N. Rengaraj, K. M. Vamsi, and G. Vijayaraghavan, "Evaluation of machine learning approaches for cirrhosis survival prediction," p. 10915362, 2025, doi: 10.1109/IITCEE64140.2025.10915362.

[56] J. Blessy R and N. A. H, "Liver cirrhosis stage classification using ensemble machine learning techniques," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 9, no. 4, p. 44060, 2025, doi: 10.55041/ijsrem44060.

[57] M. Luo, D. Yan, X. Wang, Y. Wang, H. Li *et al.*, "Construction of a machine learning prognostic prediction model based on psoas muscle index for patients with decompensated liver cirrhosis," *Chinese Journal of Hepatology*, vol. 33, no. 7, 2025, doi: 10.3760/cma.j.cn501113-20231123-00222.

[58] P. K. Mishra, B. K. Chaurasia, and M. M. Shukla, "XAIHO: explainable AI leveraging hybrid optimized framework for liver cirrhosis detection," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 470, 2025, doi: 10.1007/s44163-025-00470-y.

[59] D. Kurniawan and M. Yasir, "Optimization sentimen analysis using CRISP-DM and Naive Bayes methods implemented on social media," *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, vol. 6, no. 2, pp. 127–135, 2022, doi: 10.22373/cj.v6i2.12793.

[60] Kaggle, "Primary biliary cirrhosis (pbc) dataset," 2023, accessed: 05 Ago. 2024.

[61] J. C. G. Mejia, F. A. V. Agudelo, D. A. M. Samboní, and M. F. C. Ortega, "Method for the application of BI in software development," *Salud, Ciencia y Tecnología*, vol. 5, p. 1424, 2025, doi: 10.56294/saludcyt20251424.

[62] U. Roman-Concha, A. López, K. Ruiz-Carrasco, C. Chavez-Herrera *et al.*, "Propensity model using decision trees (LightGBM) for the management of the effective credit product in a financial entity," *Advances in Artificial Intelligence and Machine Learning*, vol. 5, no. 1, p. 51184, 2025, doi: 10.54364/AAIML.2025.51184.

[63] D. A. Hamidah, R. Salkiawati, and R. Sari, "Analisis sentimen ulasan customer kopi TMLST menggunakan algoritma Naïve Bayes," *Journal of Students' Research in Computer Science*, vol. 5, no. 1, pp. 89–97, 2024, doi: 10.31599/mrm89y71.

[64] J. Adiputra and D. Mahdiana, "Analisis Sentimen Degan Algoritma Support Vector Machine Terhadap Penyakit Hepatitis Akut Misterius," *IDEALIS: InDonEsiA journaL Information System*, vol. 6, no. 1, p. 2985, 2023, doi: 10.36080/idealis.v6i1.2985.

[65] J. de Pedro-Carracedo, J. Clemente, D. Fuentes-Jimenez, M. F. Cabrera-Umpiérrez, and A. P. Gonzalez-Marcos, "Photoplethysmographic signal-diffusive dynamics as a mental-stress physiological indicator using convolutional neural networks," *Applied Sciences*, vol. 13, no. 15, p. 8902, 2023, doi: 10.3390/app13158902.

[66] A. Fajriansyah, D. Yusup, and K. Prihandini, " Prediski Kelulusan Nilai Kalkulus Menggunakan Naive Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, p. 13343, 2025, doi: 10.36040/jati.v9i2.13343.

[67] R. M. Apsari, "Penerapan metode Naïve bayes dalam memprediksi prestasi siswa," *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 4, no. 2, pp. 76–85, 2024, doi: 10.55382/jurnalpustakaai.v3i3.760.

[68] S. K. Kamath, S. K. Pendekanti, and D. Rao, "Livmarx: An optimized low-cost predictive model using biomarkers for interpretable liver cirrhosis stage classification," *IEEE Access*, vol. 12, pp. 98 712–98 725, 2024, doi: 10.1109/ACCESS.2024.3422451.

[69] X. Jiang, R. Zhou, F. Jiang, Y. Yan, Z. Zhang, and J. Wang, "Construction of diagnostic models for the progression of hepatocellular carcinoma using machine learning," *Frontiers in Oncology*, vol. 14, p. 1401496, jun 2024, doi: 10.3389/fonc.2024.1401496.

[70] M. Tabatabai, D. Wilus, C. K. Chen, K. P. Singh, and T. L. Wallace, "Taba binary, multinomial, and ordinal regression models: New machine learning methods for classification," *Bioengineering*, vol. 12, no. 1, p. 2, jan 2025, doi: 10.3390/bioengineering12010002.

[71] S. Vithal, S. Shah, V. C. Kulkarni, and S. Terdal, "Liver cirrhosis prediction using random forest," *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS)*, vol. 14, no. 5, pp. 115–122, 2025, doi: 10.51583/ijltemas.2025.140500115.

[72] B. Da, H. Chen, W. Wu, W. Guo, A. Zhou, Q. Yin, J. Gao, J. Chen, J. Xiao, L. Wang, M. Zhang, Y. Zhuge, and F. Zhang, "Development and validation of a machine learning-based model to predict survival in patients with cirrhosis after transjugular intrahepatic portosystemic shunt," *eClinicalMedicine*, vol. 79, p. 103001, jan 2025, doi: 10.1016/j.eclinm.2024.103001.

[73] F. Yao, J. Luo, Q. Zhou, L. Wang, and Z. He, "Development and validation of a machine learning-based prediction model for hepatorenal syndrome in liver cirrhosis patients using MIMIC-IV and eICU databases," *Scientific Reports*, vol. 15, no. 1, p. 3412, feb 2025, doi: 10.1038/s41598-025-86674-9.

[74] A. M. Al Alawi, H. Al Kaabi, Z. Al Falahi, Z. Al-Naamani, and S. Al Busafi, "Machine learning-powered 28-day mortality prediction model for hospitalized patients with acute decompensation of liver cirrhosis," *Oman Medical Journal*, vol. 39, no. 3, p. e624, may 2024, doi: 10.5001/omj.2024.79.

Deep Learning for Canonical Reconstruction of Deformable Objects from Depth Images in Robotic Vision and SLAM-Aware Perception

Fahd Alhamazani
Faculty of Computing and Information Technology, Department of Computer Sciences,
Northern Border University, Rafha, Saudi Arabia

***Abstract*—Reconstructing the canonical pose of non-rigid objects from arbitrary depth observations is an important problem in robotic vision, particularly for systems that must perceive, track, and interact with deformable objects in dynamic environments. In robotics and SLAM-related perception, depth cameras are widely used to support object recognition, spatial understanding, scene mapping, and motion analysis. However, non-rigid deformation remains challenging because the same object may appear in significantly different poses, making reliable object-level representation and tracking difficult. In this study, we present a deep learning approach for reconstructing the default canonical pose of non-rigid objects from single depth images. The proposed model combines short-range and long-range feature extraction with the original depth input to capture both local geometric details and global structural information. By transforming arbitrary posed observations into a consistent canonical representation, the method supports more stable shape understanding, pose normalization, and object-level perception for robotic systems operating in real-world environments. This is particularly relevant to robotic vision tasks involving human motion analysis, deformable-object tracking, manipulation, and semantic mapping. The model is trained on synthetic human datasets and evaluated on synthetic human, real human, and animal datasets. Experimental results demonstrate improved retrieval accuracy compared with existing methods, showing that the proposed approach can generalize across different non-rigid categories and sensing conditions. These findings highlight the potential of canonical pose reconstruction as a useful component for intelligent robotic perception, depth-based scene interpretation, and SLAM-aware systems that require robust understanding of deformable objects.**

***Keywords*—Depth image; computer vision; feature representation; sensor fusion; object-level SLAM; human–robot interaction**

I. Introduction

Reconstructing the default pose of non-rigid objects from arbitrary poses has become an increasingly important task in fields such as computer vision, computer graphics, medical imaging, and robotics. Non-rigid objects, such as human bodies, animals, and deformable objects, undergo complex deformations that make accurate reconstruction and pose normalization a significant challenge. This task is crucial for numerous applications, including animation, motion tracking and shape analysis. Understanding how to reconstruct a standard or default pose from an arbitrary one is not only vital for precise shape representation but also enhances the ability of intelligent systems to interpret and manipulate 3D objects.

In many real-world applications, obtaining consistent 3D representations from varying poses is necessary. For instance, in animation, characters often need to be represented in a canonical pose to ensure uniformity across different scenes and actions. Similarly, in medical imaging, normalizing poses allows for more consistent comparisons of anatomical structures, which is crucial for accurate diagnostics and surgical planning. In robotics, understanding non-rigid body deformations, such as the movements of animals or humans, helps in designing systems that can interact with or mimic these behaviors effectively.

Existing methods for 3D shape reconstruction have focused on either rigid or non-rigid objects, with varying degrees of success. Traditional rigid body reconstruction methods, such as Principal Component Analysis (PCA) or rigid transformation matrices, are not effective for non-rigid objects, as they cannot handle the complex, localized deformations that occur in flexible bodies. On the other hand, non-rigid reconstruction methods, such as Multidimensional Scaling (MDS) [1], [2], Global Point Signatures (GPS) [3], and detail-preserving mesh unfolding techniques [4], have been designed to capture these deformations but often struggle to balance between global structure preservation and fine-grained local details. Many of these methods require tuning or additional constraints to ensure accurate results, particularly when dealing with large datasets or objects with high variability in pose and shape.

Our work introduces a novel approach for reconstructing the default pose of non-rigid objects using depth images as input. Depth images offer rich geometric information, making them a valuable input format for understanding object deformations. By leveraging both short-range and long-range features, our model can effectively capture fine-grained details, such as edges and textures, while also maintaining the overall structure of the object. This dual approach allows the model to handle complex deformations that occur across varying scales, making it suitable for reconstructing both local and global characteristics of non-rigid objects.

We focus on three main components within our model: short-range feature extraction, long-range feature extraction, and the final reconstruction component. Short-range features help capture detailed deformations in small regions of the object, while long-range features ensure that the model accounts for broader spatial relationships across the entire object. The combination of these components, along with the original input image, enables the model to reconstruct the default pose with

high accuracy. This approach is particularly effective for non-rigid bodies, where both local and global deformations must be accounted for.

In summary, this study presents a new method for reconstructing the default pose of non-rigid objects from arbitrary poses using depth images. By combining short-range and long-range feature extraction, our approach provides a comprehensive solution for capturing both fine details and global structure in non-rigid objects. The results of our experiments show that the model is highly effective in various settings, including human and animal datasets, and generalizes well even when trained on synthetic data. Our findings contribute to the growing body of research in shape analysis and 3D reconstruction, with potential applications in animation, medical imaging, and robotics.

Overall, the main contributions of this study are summarized as follows:

- A short-range feature extraction method to capture local details.
- A long-range feature extraction method to capture global structures.
- The results demonstrate that our model successfully recovers the canonical pose for unseen samples.

II. Related Work

The problem of deforming non-rigid objects to a canonical pose from single depth images has been widely addressed in the literature through various methodologies, including learning-based approaches, shape priors, and deformation networks.

One line of work focuses on Multidimensional Scaling (MDS) approaches. For example, Fast-MDS, introduced by Faloutsos and [1], was an early attempt to reduce the computational overhead associated with MDS, allowing for faster shape matching and reconstruction. Despite its efficiency, Fast-MDS struggles with complex non-rigid deformations, limiting its applicability for canonical pose reconstruction from single depth images. Similarly, Non-Metric MDS and Least Squares MDS, proposed by [2], were designed to handle surface bending and stretching, but these methods also face challenges when dealing with high variability in non-rigid shapes and often require multiple observations to achieve robust results.

Recent advancements have expanded upon these MDS-based methods. Constrained MDS, introduced by [5], integrates additional shape constraints to improve non-rigid shape matching and deformation. This method shows improvements over previous MDS techniques but is limited in its ability to handle single-depth-image inputs, as it still relies on some prior shape knowledge.

In addition to MDS-based methods, shape reconstruction techniques have evolved to focus on the preservation of local and global structures. Global Point Signatures (GPS), combined with skeleton-based matching, was proposed by [3] as an effective approach for non-rigid shape matching. GPS, however, falls short in scenarios where fine-grained deformations and high-resolution details are necessary for accurate canonicalisation.

One of the more recent approaches, Detail-preserving Mesh Unfolding also, developed by [4], focuses on preserving intricate surface details while deforming a shape to its canonical pose. This method has been particularly effective in handling high-resolution surfaces, though it requires significant computational resources, which limits its scalability to real-time applications.

Moreover, learning-based approaches like Deform2NeRF and DIF-Net [6], [7] have shown promising results in deforming objects into canonical poses. These methods leverage neural deformation fields and 2D-3D feature fusion to accurately reconstruct non-rigid objects. While these methods achieve high accuracy in reconstructing complex deformations, their heavy reliance on neural networks introduces high computational costs and data requirements, making them less practical for real-time applications or scenarios with limited training data.

Despite the advancements, the field still faces challenges in developing methods that can robustly handle highly complex, non-rigid deformations with minimal data and in real-time scenarios. Methods that balance computational efficiency with high reconstruction fidelity are crucial for advancing the state-of-the-art in non-rigid shape reconstruction and canonical pose estimation.

III. Methodology

The goal is to restore a non-rigid body to its default pose. The shape, represented in a 2.5D image, may appear in any pose and could be subject to significant deformation, causing parts of the shape to become occluded or heavily distorted. Therefore, the model is designed to learn how to reconstruct the non-rigid shape in a canonical pose, The complete model is illustrated in Fig. 1.

The model consists of three components. The first stage, Section III-A, captures fine-grained, localized information (e.g., edges), which helps in accurately modeling small-scale deformations in parts of the object. This provides the model with crucial details on how local areas deform during movement, essential for recovering fine details in the reconstructed default pose.

The second component, Section III-B, captures spatial relationships over larger distances within the object. These features ensure that parts of the object that are far apart, such as limbs in a human body, remain correctly aligned relative to each other. This step helps maintain the global coherence of the object's structure during reconstruction, ensuring the reconstructed pose is anatomically accurate and free from unnatural distortions.

Finally, the third stage, Section III-C processes the outputs of both the first and second stages to reconstruct the shape in its default pose.

A. Short Range Features

Extracting short range features features from depth images, especially in the context of non-rigid shape deformation help model to focus on the fine-grained information in small regions of the image, such as edges, textures, or surface variations [8]. This helps in identifying detailed characteristics of the shape,

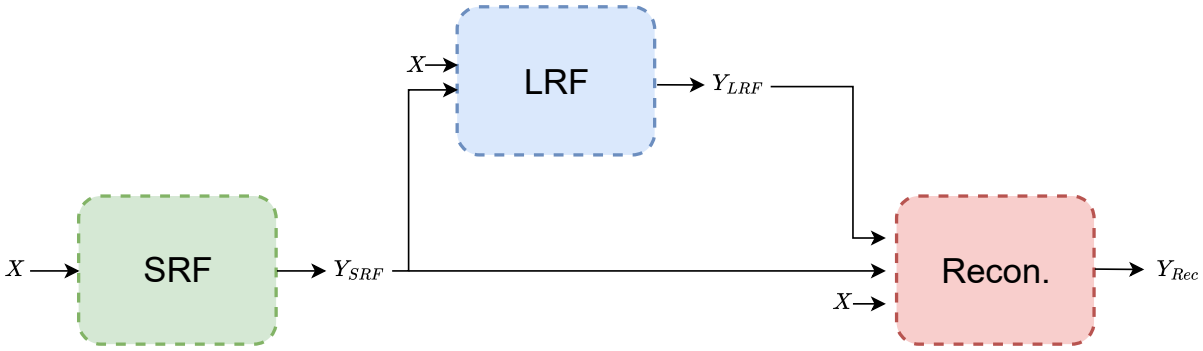


Fig. 1. The complete model first extracts short-range features. Afterward, both the original input and the output of the SRF model (Y_{SRF}) are used as inputs for the long-range features model, which extracts scaled features (Y_{LRF}). Finally, the canonical pose is reconstructed through the reconstruction model (Y_{Rec}).

which are essential for accurately representing and reconstructing the object's geometry in its default pose. Furthermore, Non-rigid objects can deform in complex ways (bending, stretching, etc.). Local features can capture these deformations at small scales, which helps in tracking and understanding how different parts of the shape move or change during deformation (see Fig. 2).

Given a depth image I, where $I \in \mathbb{R}^{500\times500}$, the Short Range Features (SRF) consist of five convolution layers that process the input depth image to extract the short range features:

$$Y_{SRF} = SRF(I)$$

SRE consist of N block, where $N = 9$, each block consist of convolution and Transpose-convolution. At the end, the output Y_{SRF} going to be $Y_{SRF} \in \mathbb{R}^{500\times500}$.

B. Long Range Features

Capture the broader spatial dependencies within the depth image and are crucial for maintaining the overall structure and coherence of the non-rigid object during reconstruction. These features help the model account for how distant parts of the object relate to one another, ensuring that even large-scale deformations are represented accurately. The concatenation of the short-range features with the input image provides a richer representation, enabling the model to leverage both local, fine-grained details and broader spatial information. This holistic understanding ensures that distant parts of the object, such as the extended surfaces, are correctly aligned and positioned relative to one another in the default pose [9] [8].

$$Y_{LRF} = LRF(I, Y_{SRF})$$

Given I and Y_{SRF} input depth image and extracted local features, respectively, the Long Range Feature (LRF) process both the image and the extracted features to produce the long features. The model consist of N block where $N = 5$, each block consist of convolution layer with dilation value to capture long range. The dilation range values in order $dilation = [2, 3, 4, 5, 6]$.

C. Default Pose Reconstruction

The reconstruction component synthesizes all this information to output the default pose in the form of a depth image. The model uses both localized details (from short-range features) and overall structure (from long-range features) and to infer the original, undeformed pose (input depth image) were also utilised. This approach allows the Default Pose Reconstruction (DPR) component to handle complex deformations by integrating both fine-grained and large-scale spatial information, ensuring accurate and detailed default pose reconstruction.

$$Y_{DPR} = DPR(I, Y_{SRF}, Y_{LRF})$$

Given I, Y_{SRF}, and Y_{LRF}, the DPR component processes this information to reconstruct a shape in the default pose, Y_{DPR}, where $Y_{DPR} \in \mathbb{R}^{500\times500}$. The component consists of an Encoder-Decoder architecture, where the encoder is made up of $N = 4$ blocks, and the decoder consists of $K = 4$ blocks. Each encoder block contains convolution, LeakyReLU activation, and max-pooling layers to reduce dimensionality, while each decoder block employs transposed convolution and ReLU activation.

D. Loss Function

The model utilise two loss functions (see Fig. 3):

Depth Loss. We employ the Mean Squared Error for the depth loss:

$$L_{Depth} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_d - y_d)^2$$

Here, $\hat{y}_d$ and y_d represent the predicted depth and the ground truth depth, respectively.

Mask Loss. As we want the model focus on the shape only, we make the model learn the mask of the shape by adding mask loss:

$$L_{Mask} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_m - y_m)^2$$

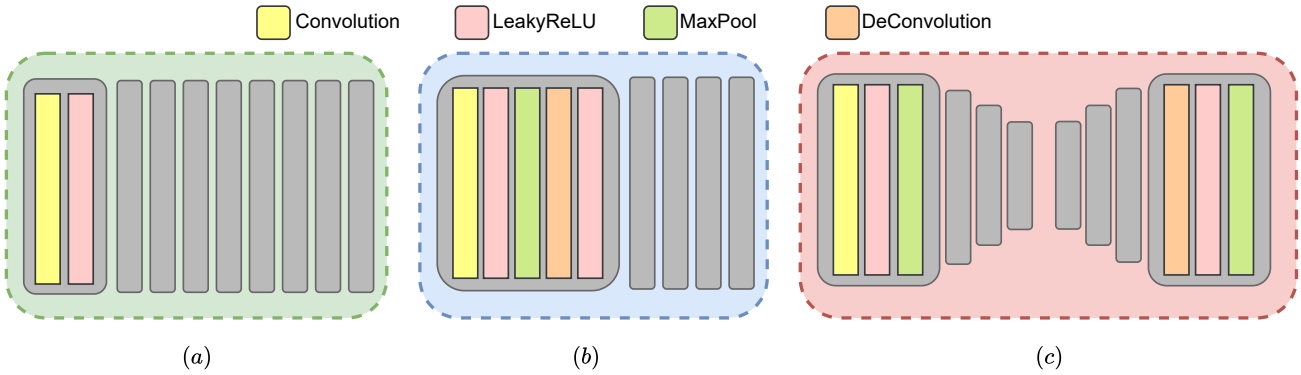


Fig. 2. (a) Shows the SRF layers, where the sub-model consists of nine blocks. Each block contains a convolution layer followed by a LeakyReLU activation. (b) Shows the LRF layers, consisting of five blocks, where each block includes both convolution and deconvolution operations with different dilation rates. (c) Shows the reconstruction sub-model, which includes an encoder–decoder architecture.

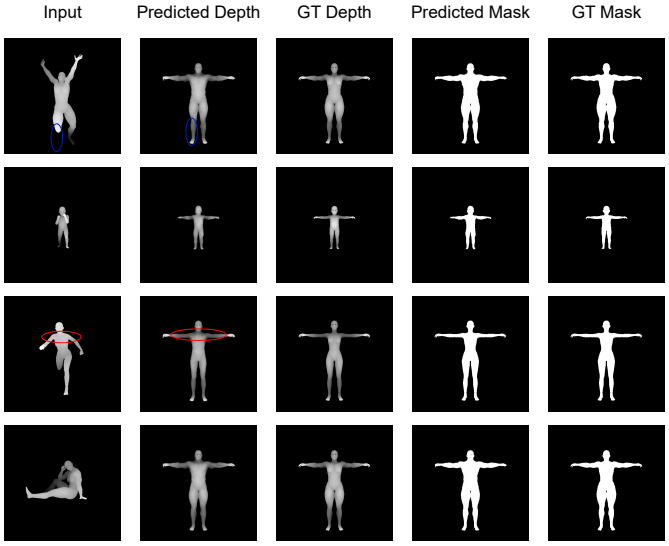


Fig. 3. Our model results on the synthetic dataset. The model predicts both the depth image (second column) and the mask (fourth column). The blue circle shows that the model can recover missing parts. The red circle shows that the model can correctly predict the depth values for the shoulders.

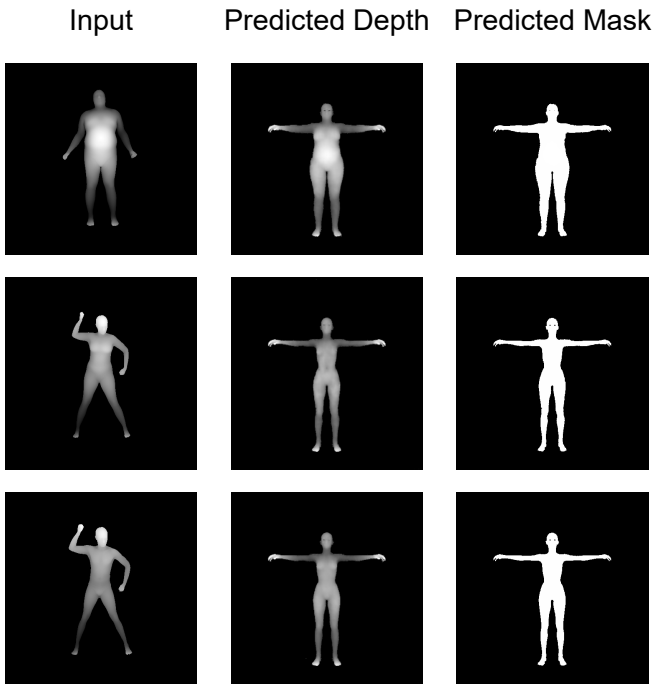


Fig. 4. Our model results on real human datasets. The model was trained on synthetic data and then tested on a real human dataset.

Combined Loss. We introduce coefficients α and β to balance the training:

$$L_{weighted} = \alpha L_{Depth} + \beta L_{Mask}$$

IV. Experiments

A. Training Details

We trained our model for 500 epochs using an end-to-end approach. In the initial phase of training, we set the weighting parameter $\beta = 1$ and $\alpha = 0.01$ for the first 100 epochs. This strategy was employed to allow the model to primarily focus on learning the mask of the shapes during the early stages, thereby reducing the overall complexity of learning the canonical form in the depth format. After the first 100 epochs, both α and β were set to 1, enabling the model to simultaneously learn both the mask and the canonical form with equal emphasis. This progressive learning approach was validated through an ablation study, which demonstrated its effectiveness in simplifying the learning process for depth-based reconstruction.

The model was trained using the Adam optimizer with a learning rate of 0.001, on an NVIDIA RTX 4090 GPU with 24 GB of memory. The total training time was approximately three days with a batch size of 8.

B. Dataset

We conducted our experiments using three datasets: two human datasets and one animal dataset.

The first dataset, provided by [10], consists of 400 samples captured from real human subjects. This dataset represents 40 unique individuals, with each subject contributing multiple samples.

The second dataset, also from [10], comprises synthetic human subjects and includes 300 samples in total. These synthetic samples offer controlled variations that complement the real-world dataset.

To demonstrate the generalizability of our model beyond human subjects, we employed the TOSCA dataset [11], which contains 3D models of animals. This allows us to evaluate the model's ability to reconstruct non-human shapes effectively.

All datasets were processed using Blender and automated with Python scripts. For each sample, we generated depth images of size 500×500, ensuring that the shapes were centered in the frame before capturing the images.

C. Evaluation

For evaluation, we followed the methodology used in previous studies [12], [13], and assessed our model based on retrieval results. Specifically, we employed the Clock Matching and Bag-of-Features (CM-BOF) approach to produce the retrieval results. For a more detailed explanation of CM-BOF, refer to [14].

We selected four evaluation metrics to assess the performance of our model. The first metric is Nearest Neighbour (NN), which finds the nearest neighbor to a point using a specified distance metric. The second metric is First Tier (FT), which measures the precision at the first rank or within the top-n results of the retrieval. The third, Second Tier (ST), extends FT by considering a larger set of result (top-m where $m > n$). Finally, we use Discounted Cumulative Gain (DCG) to evaluate the effectiveness of the ranking algorithm by taking into account the relevance of retrieved items and their positions in the ranking.

D. Comparison to Prior Works

For comparison, we selected several methods from previous studies. These include the Fast-MDS approach proposed by [1], and the Non-Metric MDS method, based on Multidimensional Scaling (MDS), introduced by [2]. Additionally, we included Least Squares MDS from the same study [2].

We also compared our method against the Accelerated MDS approach [15], which improves the efficiency of traditional MDS algorithms. Furthermore, we considered Constrained MDS, a method that integrates specific shape constraints during the MDS process [5].

In addition to MDS-based methods, we evaluated our approach against shape analysis techniques such as Global Point Signatures (GPS), and a Skeleton-based method [3], both of which have demonstrated effectiveness in shape matching tasks. Finally, we included the Detail-preserving Mesh Unfolding method [4], which focuses on preserving fine geometric details during shape reconstruction.

TABLE I. RETRIEVAL RESULTS FOR SYNTHETIC HUMAN DATASET

	NN	FT	ST	DCG
Classic MDS	0.10	0.22	0.39	0.54
Fast MDS	0.14	0.20	0.35	0.53
Non-metric MDS	0.09	0.24	0.41	0.55
Least Square MDS	0.01	0.13	0.31	0.45
Constrained MDS	0.04	0.14	0.25	0.46
GPS	0.40	0.20	0.32	0.56
Mesh Unfolding	0.04	0.18	0.34	0.49
Skeleton-based	0.01	0.14	0.32	0.46
Our	**0.61**	**0.53**	**0.73**	**0.78**

TABLE II. RETRIEVAL RESULTS FOR REAL HUMAN DATASET, TRAINED ON SYNTHETIC HUMAN DATASET AND TESTED ON REAL HUMAN DATASET

	NN	FT	ST	DCG
Classic MDS	0.01	0.03	0.07	0.28
Fast MDS	0.00	0.02	0.04	0.27
Non-metric MDS	0.02	0.04	0.08	0.30
Least Square MDS	0.00	0.00	0.01	0.26
Constrained MDS	0.00	0.01	0.03	0.27
GPS	0.07	0.06	**0.12**	0.33
Mesh Unfolding	0.00	0.01	0.03	0.28
Skeleton-based	0.01	0.01	0.02	0.27
Our	**0.1**	**0.09**	0.1	**0.41**

E. Result

We evaluated our model on three datasets: synthetic humans, real humans, and animals. For the synthetic human dataset, we split the subjects into a training set of 10 subjects and a test set of 5 subjects (the results shown in Table I). Our model demonstrated a noticeable improvement over previous methods, showing a significant margin of enhancement in the retrieval results.

For the real human dataset (the results shown in Table II and Fig. 4), we followed an unseen setup, where the model was trained on the synthetic dataset and tested on the real human data. Despite the challenge of domain adaptation, our model exhibited substantial improvements, further confirming its generalization capabilities.

Finally, for the animal dataset (the results shown in Table III and Fig. 5), we used the cat and horse samples as the test set, with the remaining animal data used for training. In this case, too, our model outperformed prior approaches, showing marked improvements in retrieval performance compared to existing methods.

V. ABLATION STUDIES

We conducted several experiments to evaluate the effectiveness of different components of our model. The ablation study is designed to demonstrate the impact of each component on the overall performance, the results are shown in Table IV.

A. Disabling Short-Range Features

In this experiment, we disabled the short-range feature extraction component of the model. As a result, the retrieval performance degraded significantly, indicating that short-range features are crucial for capturing fine-grained details necessary for accurate shape reconstruction.

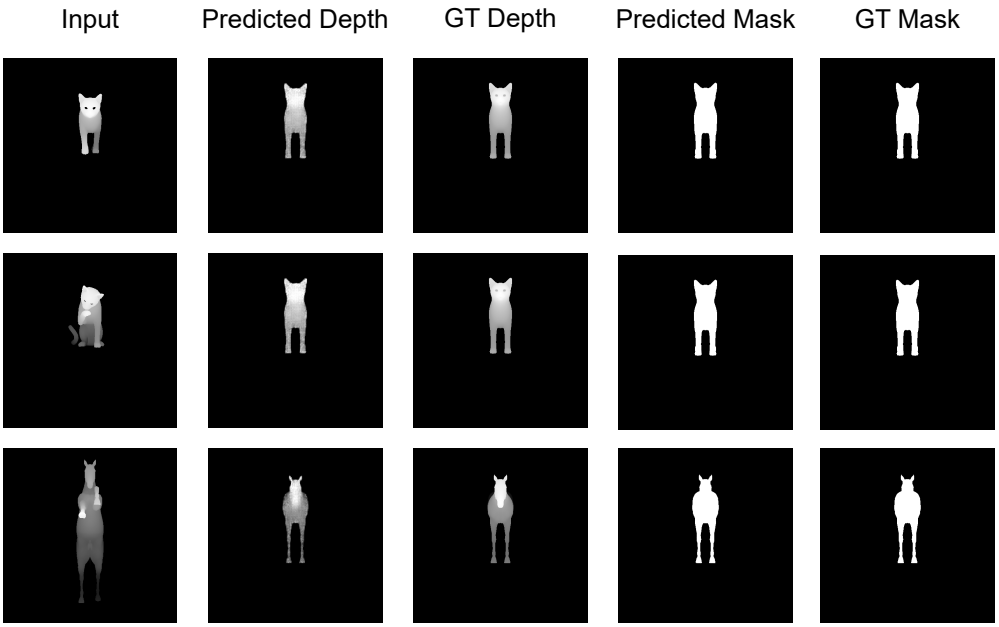


Fig. 5. Our model results on the TOSCA dataset.

TABLE III. RETRIEVAL RESULTS FOR TOSCA DATASET

	NN	FT	ST	DCG
Classic MDS	0.74	0.54	0.80	0.80
Fast MDS	0.73	0.52	0.77	0.77
Non-metric MDS	0.76	0.67	0.87	0.85
Least Square MDS	0.79	0.63	0.86	0.84
Constrained MDS	0.88	0.71	**0.89**	0.89
GPS	0.71	0.52	0.72	0.76
Mesh Unfolding	0.88	0.65	0.86	0.85
Skeleton-based	0.78	0.62	0.85	0.84
Our	**0.97**	**0.82**	**0.89**	**0.93**

TABLE IV. RETRIEVAL RESULTS FOR SYNTHETIC DATASET

	NN	FT	ST	DCG
strict train	0.56	0.49	0.66	0.73
w/o SRF	0.58	0.51	0.68	0.76
w/o LRF	0.56	0.50	0.64	0.71
Our	**0.61**	**0.53**	**0.73**	**0.78**

B. Disabling Long-Range Features

When we disabled the long-range feature extraction component, we observed a noticeable degradation in performance compared to the full model. This highlights the importance of long-range features in learning the global structure of the object and suggests that the model effectively utilizes these features to improve reconstruction accuracy.

C. Strict Train-Test Split

To further challenge the model, we performed a strict split, where both pose and subject were excluded from the training set. We implemented this as a cross-validation process, ensuring that the model was tested on all poses and subjects. The model showed resilience under this challenging setup, further validating its generalization ability.

VI. CONCLUSION

In this work, we presented a model for reconstructing the default pose of non-rigid objects from any given pose using depth images as input. Through a combination of short-range and long-range feature extraction, along with the original input, our model demonstrated significant improvements over previous methods, particularly in complex shape deformation scenarios.

Our experiments on synthetic and real human datasets, as well as an animal dataset, showed the model's robustness and generalization capabilities. Specifically, our model achieved notable gains in retrieval accuracy, especially in challenging setups, where the training and test data included unseen poses and subjects.

REFERENCES

[1] C. Faloutsos and K.-I. Lin, "FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," in *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, 1995, pp. 163–174.

[2] A. Elad and R. Kimmel, "On bending invariant signatures for surfaces," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 10, pp. 1285–1295, 2003.

[3] D. Pickup, X. Sun, P. L. Rosin, and R. R. Martin, "Skeleton-based canonical forms for non-rigid 3D shape retrieval," *Computational visual media*, vol. 2, pp. 231–243, 2016.

[4] Y. Sahillioğlu and L. Kavan, "Detail-preserving mesh unfolding for non-rigid shape retrieval," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, pp. 1–11, 2016.

[5] Y. Sahillioğlu, "A shape deformation algorithm for constrained multidimensional scaling," *Computers & Graphics*, vol. 53, pp. 156–165, 2015.

[6] Y. Deng, J. Yang, and X. Tong, "Deformed implicit field: Modeling 3d shapes with learned dense correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 286–10 296.

[7] X. Xie, X. Guo, W. Li, J. Liu, and J. Xu, "Deform2nerf: Non-rigid deformation and 2d–3d feature fusion with cross-attention for dynamic human reconstruction," *Electronics*, vol. 12, no. 21, p. 4382, 2023.

[8] D. Liu, X. Liu, and Y. Wu, "Depth reconstruction from single images using a convolutional neural network and a condition random field model," *Sensors*, vol. 18, no. 5, p. 1318, 2018.

[9] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 297–306, 2019.

[10] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. B. Hamza, A. Bronstein, M. Bronstein *et al.*, "Shape retrieval of non-rigid 3d human models," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 169–193, 2016.

[11] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

[12] D. Pickup, J. Liu, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, S. Nie, L. Jin, G. Shamai *et al.*, "An evaluation of canonical forms for non-rigid 3D shape retrieval," *Graphical Models*, vol. 97, pp. 17–29, 2018.

[13] A. Bronstein, M. Bronstein, U. Castellani, B. Falcidieno, A. Fusiello, A. Godil, L. Guibas, I. Kokkinos, Z. Lian, M. Ovsjanikov *et al.*, "SHREC 2010: robust large-scale shape retrieval benchmark," *Proc. 3DOR*, vol. 5, no. 4, pp. 1–8, 2010.

[14] Z. Lian, A. Godil, X. Sun, and J. Xiao, "CM-BOF: visual similarity-based 3D shape retrieval using clock matching and bag-of-features," *Machine Vision and Applications*, vol. 24, pp. 1685–1704, 2013.

[15] G. Shamai, M. Zibulevsky, and R. Kimmel, "Accelerating the computation of canonical forms for 3D nonrigid objects using multidimensional scaling," in *Proceedings of the 2015 Eurographics Workshop on 3D Object Retrieval*, 2015, pp. 71–78.

Accelerating Blockchain Consensus: A Parallel Mining Approach for High-Throughput, Low-Latency Networks

Sohail Jabbar
College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh, 11432, Saudi Arabia

Abstract*—*Problem: **Blockchain consensus remains constrained by slow transaction verification, redundant mining effort, high communication overhead, and increased confirmation latency as the number of nodes grows.** ***Objective:*** **This study proposes NCABS, a controlled parallel mining consensus approach intended to improve blockchain throughput, latency, transaction commit rate, and resource utilization.** ***Methodology:*** **NCABS combines verifiable head-miner selection, transaction-validation delegation, parallel block generation, and a 75% confirmation threshold. The framework is evaluated against Practical Byzantine Fault Tolerance (PBFT) through ten repeated experiments over node densities from 40 to 140, using throughput, latency, communication load, transaction commit rate, scalability, and responsiveness as performance metrics.** ***Results:*** **The plotted results show that NCABS achieves higher throughput and transaction commit rates than PBFT, reduces aggregate communication messages by approximately 45%, and lowers latency at 40 nodes from about 310 ms for PBFT to about 195 ms for NCABS. In the transaction-commit-rate experiment, NCABS reaches approximately 3200 transactions per second, compared with about 2200 transactions per second for PBFT.** ***Conclusion:*** **The findings indicate that controlled parallel mining can reduce redundant computation and communication while improving blockchain responsiveness. The current evaluation is performance-oriented; adversarial testing and broader empirical comparison with additional consensus baselines are identified as future work.**

Keywords*—*Blockchain; consensus algorithm; parallel mining; PBFT; high transaction rate; low-latency

I. Introduction

Blockchain-based systems combine cryptography, public key infrastructure, and economic modeling to synchronize distributed databases through peer-to-peer networking and decentralized consensus. This technology has the potential to revolutionize various business areas and is expected to be rapidly adopted into mainstream use in a wide range of domains [1]. Blockchain technology has immense potential, but its market penetration is still relatively low. The ecosystem surrounding blockchain is not yet mature enough to support its widespread use, especially in supply chain applications. This is because most supply chains today involve many geographically dispersed and professionally diverse stakeholders, making them complex and difficult to manage. Therefore, extensive groundwork is required before blockchain can be adopted on a large scale. A recent study by Jabbar *et al.* [2] provides a detailed analysis of blockchain-enabled supply chains, discussing the challenges and future directions. The most significant challenge the blockchain ecosystem faces is scalability, which affects transaction processing speed and efficiency. Compared to traditional databases, blockchain is slower in retrieving and committing records, and it requires more computing resources. The scalability of these resources is a major concern as it affects the system's ability to respond and function effectively as the size of input increases to meet user demand [3]. Scalability is crucial in three key aspects of a supply chain system: 1) Throughput, which refers to the efficiency of producing results; 2) Storage, which involves maintaining transaction records; and 3) Network communication, which pertains to the number of messages exchanged among nodes to finalize a transaction. This study focuses on throughput and network communication scalability.

Considering the founding cryptocurrency, Bitcoin can handle $4--5$ transactions per second. As the average transaction size is 380.04 bytes and the block size is $1,048,576$ bytes, the number of transactions per block is $2,759.12$. The current block generation time is 600 seconds (10 minutes). Conversely, traditional centralized payment systems, e.g., VISA, MasterCard, and PayPal, offer significantly higher transaction rates per second [4]. VISA processes over a million daily transactions, i.e., around 1800 per second. From a storage perspective, the size of the Bitcoin Blockchain has grown steadily since its creation. It reached approximately 242.39 gigabytes in size as in Q3, 2021, while it was only $53,647$ MB in Q4, 2015. In Ethereum, transactions increased from just USD $ 3,000 in Oct 2015 to over one million in Jan 2023. Improving network communication is crucial for reducing energy consumption and propagation delay. In Blockchain, each node acts as a relay, broadcasting all transactions at least twice. When a transaction is generated, it is first broadcast to all nodes. Once a block containing the transaction is mined, it is rebroadcast to all nodes. Unfortunately, this process leads to block propagation delay [5]. Given this process, increasing the data transmission rate to handle a large volume of transactions is not feasible, as it requires more network bandwidth. Hence, it is essential to develop more effective data transmission mechanisms. Additionally, the number of transactions per block, N_{tx}, is determined by the maximum block size, B_{size}, and the average transaction size, T_{size}, as shown in Eq. (1):

$$N_{tx} = \left\lfloor \frac{B_{size}}{T_{size}} \right\rfloor \tag{1}$$

The throughput, θ, is the rate at which transactions are processed, which is defined by the number of transactions per block and the block generation time, T_{block}, as described in Eq. (2):

$$\Theta = \frac{N_{tx}}{T_{block}} = \frac{\left\lfloor \frac{B_{size}}{T_{size}} \right\rfloor}{T_{block}} \tag{2}$$

The delay in block propagation, Δt_{prop}, is a function of the network topology and the number of nodes N, as shown in Eq. (3):

$$\Delta t_{prop} = f(N, networktopology) \tag{3}$$

The total transaction confirmation time, $T_{confirmation}$, is the sum of the transaction propagation delay, the mining time, and the block propagation delay, as defined in Eq. (4):

$$T_{confirmation} = \Delta t_{txProp} + T_{mining} + \Delta t_{blockprop} \tag{4}$$

In a parallel mining system like NCABS, the total consensus time, T_{NCABS}, is the minimum time among all parallel tasks, as shown in Eq. (5):

$$T_{NCABS} = \min(T_1, T_2, \ldots, T_k) \tag{5}$$

The communication complexity of the PBFT algorithm, C_{PBFT} is of the order of the square of the number of nodes, as shown in Eq. (6):

$$C_{PBFT} = O(N^2) \tag{6}$$

The contributions of this research study in the blockchain scalability domain are enumerated as follows. The first two are considered major contributions, while the next two are of second and third-level need and importance.

- A new consensus mechanism, NCABS, has been proposed to speed up consensus in blockchain. This mechanism uses parallel mining instead of solo or repetitive mining, making it a strong solution for achieving the target. To help readers better understand this, a graphical presentation of the various steps involved in consensus-based scalability solutions for blockchain has been included.
- A detailed experimental comparison of PBFT and NCABS is presented, focusing on throughput, network latency, scalability, and responsiveness, with results depicted.
- A comparative analysis of PBFT variants, including Multi-layer PBFT, NBFT, CDBFT, and RBFT, is presented in a summarized tabular format and a discussion.
- A multi-layer categorization of available scalability solutions in the literature is provided, along with their extended versions.

II. Problem Formulation

This study considers a blockchain network consisting of N nodes, where each node participates in transaction validation and block generation. The key challenge is to minimize:

- Transaction confirmation time ($T_{\text{confirmation}}$)
- Resource utilization (CPU, energy)
- Network communication overhead

Subject to the following constraints:

- Maintaining consensus correctness
- Avoiding forks
- Ensuring fairness among miners

Formally, the objective function is defined as Eq. (7):

$$\begin{aligned} \min \quad & T_{\text{confirmation}} + \lambda_1 \cdot \text{ResourceUsage} \\ & + \lambda_2 \cdot \text{CommunicationCost} \end{aligned} \tag{7}$$

Such that [see Eq. (8) and Eq. (9)]:

$$\text{Consensus Validity} \geq 75\% \tag{8}$$

$$\text{Fork Probability} \rightarrow \min \tag{9}$$

The formulation reflects three observations that guide the design of NCABS. First, confirmation delay increases when many miners repeat the same transaction-verification and nonce-generation work. Second, PBFT-style consensus becomes communication intensive as the network size grows because message exchange increases with node density. Third, a controlled validator and head-miner selection process is needed so that performance gains do not come at the expense of consensus correctness or fork resistance.

III. Literature Survey

Tian Junfeng *et al.* [6] proposes a parallel mining-based enhancement to blockchain consensus that improves throughput and reduces latency by enabling concurrent mining operations. It introduces coordination mechanisms to maintain consistency and security while minimizing forks. Experimental results demonstrate significant performance gains, making the approach suitable for scalable, high-performance blockchain applications.

Yin Lingyuanan *et al.* [7] proposes a parallelism-enabled blockchain consensus framework to overcome the limitations of traditional PoW-based systems. By allowing concurrent mining and validation processes, the approach improves throughput and reduces latency, while maintaining consistency and security through coordination mechanisms. Experimental results confirm its effectiveness for scalable and real-time blockchain applications. The work by Lulu Ke *et al.* [8] proposes a blockchain-based framework for vaccine supply management that ensures verifiability, traceability, and data integrity. By eliminating reliance on centralized authorities and introducing an improved digital signature scheme, the

approach enhances both security and efficiency. Experimental results demonstrate reduced computational cost and improved trust, making it suitable for real-world healthcare supply chains.

Soohyeong Kim *et al.* [9] classified the scalability methods as on-chain, off-chain, side-chain, child-chain, and inter-chain solutions. Junfeng Xie *et al.* [10] categorized scalability solutions into four groups: the number of transactions, block interval time, data storage, and data transmission. This study classifies blockchain scalability solutions into three categories: on-chain, off-chain, and consensus mechanism-based. The following paragraphs briefly review each category, and Fig. 1 provides a visual representation of the classification. Although several solutions exist in each category, only representative methods are discussed due to space constraints.

The statistical gap motivating scalability research is substantial. Bitcoin typically processes about 4–5 transactions per second, while Ethereum is commonly reported at around 15 transactions per second and permissioned systems can achieve higher rates under controlled membership. In contrast, conventional payment systems such as VISA are reported to process around 1800 transactions per second in the cited literature [4]. This gap is intensified by communication complexity: PBFT-based designs require all-to-all message exchange with $O(N^2)$ communication cost [11], whereas scalable blockchain deployments require transaction throughput and confirmation latency to remain stable as the number of nodes increases.

A. On-Chain Solutions

Require a structural or fundamental change to the blockchain, and thus a modification to the protocol's underlying rules. This is technically known as a *hard fork* or *controversial hard fork* when there is a split in the community, as groups form who accept or reject the proposed update. Some popular examples of new currencies addressing the scalability issue that have resulted from a hard fork are as follows: *Bitcoin Cash* - Increases the block limit $(1MB - 8MB)$, *Litecoin* - Reduces block generation time $(10min - 2.5min)$, *DASH* - Higher transaction speed. For comparison, in the case of Ethereum, the maximum capacity is 15 transactions per second, and permissioned Blockchains typically offer much higher transaction throughput than an open and permissionless network. Changes to the Bitcoin Blockchain's proof-of-work parameters could be an option to increase scalability; for example, decreasing the block generation time or increasing the block size [12]. Still, these are unlikely to be feasible solutions, since they cause effects such as orphan blocks, forks, chain reorganization, slower network propagation, and, in extreme cases, security vulnerabilities such as selfish mining and double-spend attacks. Another method in this category to improve Blockchain scalability is *sharding*. This is a way to divide blockchain data across nodes, enabling transactions to be processed in parallel. Instead of every node processing every transaction and storing all the data, nodes are grouped, with each group handling and processing a portion of the data. This helps scale the network by enabling high concurrency. *Segwit* (Segregated Witness) is an upgrade to the data structure (transaction format) of the Bitcoin protocol. The transaction is segregated into original data and signature data (witness). The witness data is separated from the Merkle tree record [13] and appended as a separate structure at the end. This upgrade is part of the Litecoin protocol.

B. Off-Chain Solutions

It uses secondary protocols built on top of the main blockchain and are, therefore, referred to as a *second-layer* scalability solutions. In this approach, transactions are offloaded from the main blockchain and performed privately between the interacting partners. It confers the advantages of reduced congestion on the MainNet, increased throughput, reduced transaction fees, and space savings. Many off-chain solutions for various platforms are available in the literature, e.g., [14], [15]. *Lightning Network* [14] covers small and everyday transactions that do not need to be stored on the Bitcoin MainNet. Off-chain approaches to scalability issues in Ethereum include the Raiden network and Plasma cash [16], [17]. These off-chain mechanisms reduce the number of transactions of the entire network processes, though they still require at least two transactions to initialize and terminate the off-chain channels.

Optimizing the working of the ***Consensus algorithm*** can also aid in addressing scalability [15] issues. The **Proof of Work** (PoW) was first introduced with Bitcoin. It engages all miners simultaneously in a competition to create a block, and only the first to produce a valid block is accepted and rewarded. Thus, the energy and computational resources of the other nodes are wasted [15]. In contrast to PoW, **Proof of Stake** (PoS) chooses a validator among all the participating nodes using a stake-based selection algorithm. An exclusive right to create a block is assigned to this selected node, while the remaining nodes do not need to waste energy or computational resources. In contrast to PoS, **Delegated Proof of Stake** (DPoS) operates democratically, with delegates elected by token-holding nodes. EOS.IO is a decentralized Operating System that uses DPoS as its consensus algorithm, focused on scalability and capable of processing millions of transactions per second. *Steemit* featuring STEEM, SteemPower, and SteemDollar as native cryptocurrencies is a Blockchain-based social media platform. It uses DPoS and Graphene to process millions of transactions per second efficiently. In the domain of financial transactions, *Bitshare* is a decentralized platform designed to enhance transaction processing speed. Another DPoS-based solution to the scalability issue in blockchain at the second layer is *Lisk*.

It is a modular blockchain platform with smart contract functionality, which offers a partial solution to the blockchain scalability problem through the use of sidechains. Until now, all the extensions mentioned here have not reached maturity and are still under development. The **Byzantine Fault Tolerance** (BFT) consensus algorithm gives a solution to a classic problem in distributed computing, the *Byzantine General Problem*. BFT takes different forms: practical BFT, Federated Byzantine Agreement, Delegated BFT, and others. Practical BFT (PBFT) can scale to large computational loads with only a minimal increase in latency [11]. Among many other projects that use the BFT are LibraBFT, Hyperledger Fabric, and Zilliqa. The *Stellar* Consensus Protocol (SCP) is an improved implementation of the Federated Byzantine Agreement. It requires multiple rounds of voting to reach

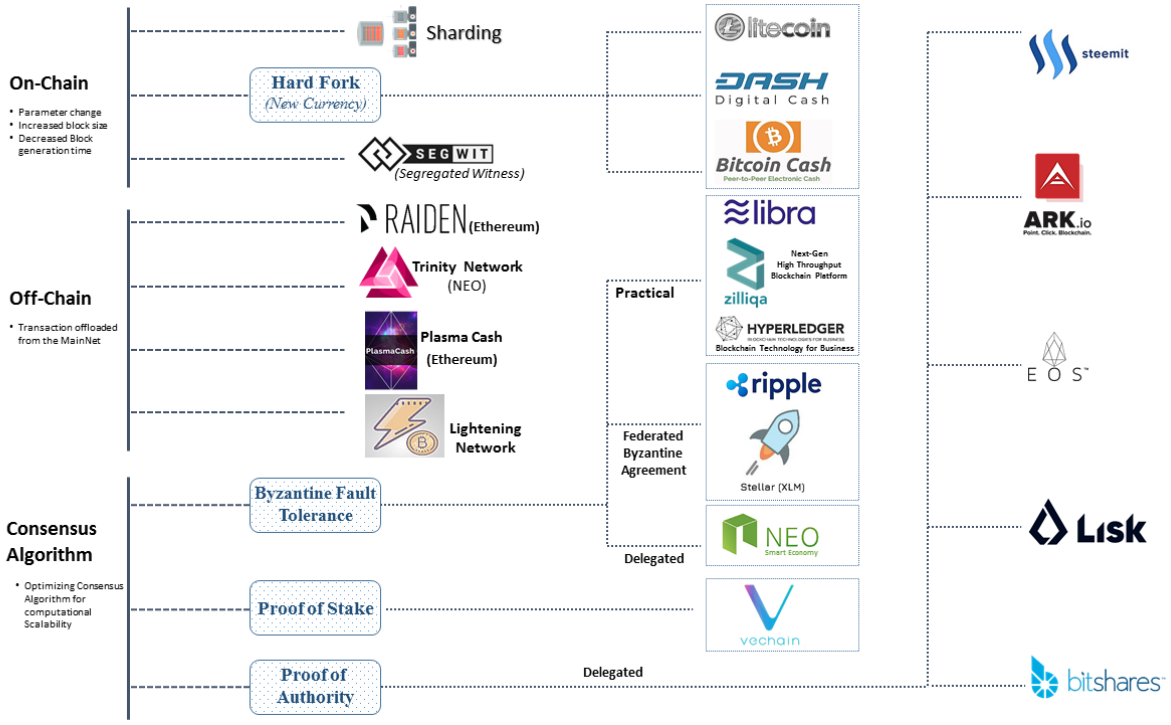


Fig. 1. Categorization of scalability solutions for the blockchain.

consensus, which generates significant network traffic. However, the messages are lightweight and within the capabilities of the currently available Internet. A further extension to SCP is the *Pi* Consensus algorithm, specifically designed for low-energy/computationally power devices such as home PCs and mobile phones. The *Bitcoin-NG* consensus algorithm [18] reduces latency and increases throughput by using leader election to append microblocks.

Some other extensions of PBFT that work for the scalability of the Blockchain network are Kauri [19], Hotstuff [20], and DBFT [21]. A comprehensive performance modeling and analysis of Hotstuff for Blockchain consensus is given in [22]. A detailed article on the challenges of PBFT-inspired consensus for blockchain and enhancements over neo dBFT is available at [23]. This proposed scalability solution, NCABS, is categorized as optimizing the Consensus algorithm to improve the scalability of the Blockchain network. Existing methods either improve throughput at the cost of communication overhead or reduce energy consumption while sacrificing decentralization. NCABS addresses both issues, simultaneously. Table I presents the comparison of blockchain consensus methods.

TABLE I. COMPARISON OF BLOCKCHAIN CONSENSUS METHODS

Method	Parallelization	Limitation	Gap
PBFT	No	$O(N^2)$ communication	Not scalable
PoW	No	Energy waste	Inefficient
Parallel Mining (existing)	Yes	Fork risk	No control
NCABS (Proposed)	Yes	—	Controlled & efficient

IV. PRELIMINARIES

The following is a brief description of key terminology and concepts to better understand the proposed solution and the rest of the study.

A. Cryptocurrency Addressing

An alphanumeric identifier of 26-35 characters representing different cryptocurrency presentation formats, beginning with the numbers 1, 3, or bc1, is shown in Fig. 2. The figure provides a self-explanatory depiction of the formats, characteristics, validation, and generation of cryptocurrencies.

The public key (K_pub) is first hashed using SHA-256 and then RIPEMD-160 to produce a public key hash $(Hpub)$.

$$H_{pub} = \text{RIPEMD160}(\text{SHA256}(K_{pub})) \tag{10}$$

The resulting public key hash is shown in Eq. (10). Next, a version byte (V) is prepended to the hash to create the versioned payload (P_V):

$$P_V = V \mid H_{pub} \tag{11}$$

Eq. (11) illustrates the concatenation of the version byte with the hash. A 4-byte checksum (C) is generated by performing a double-SHA-256 hash on the payload and extracting the first four bytes.

$$C = \text{Substr}_{0,4}(\text{SHA256}(\text{SHA256}(P_V))) \tag{12}$$

The checksum calculation is formally defined in Eq. (12). The checksum is appended to the versioned payload to form the complete binary address (A_bin).

$$A_{bin} = P_V \mid C \tag{13}$$

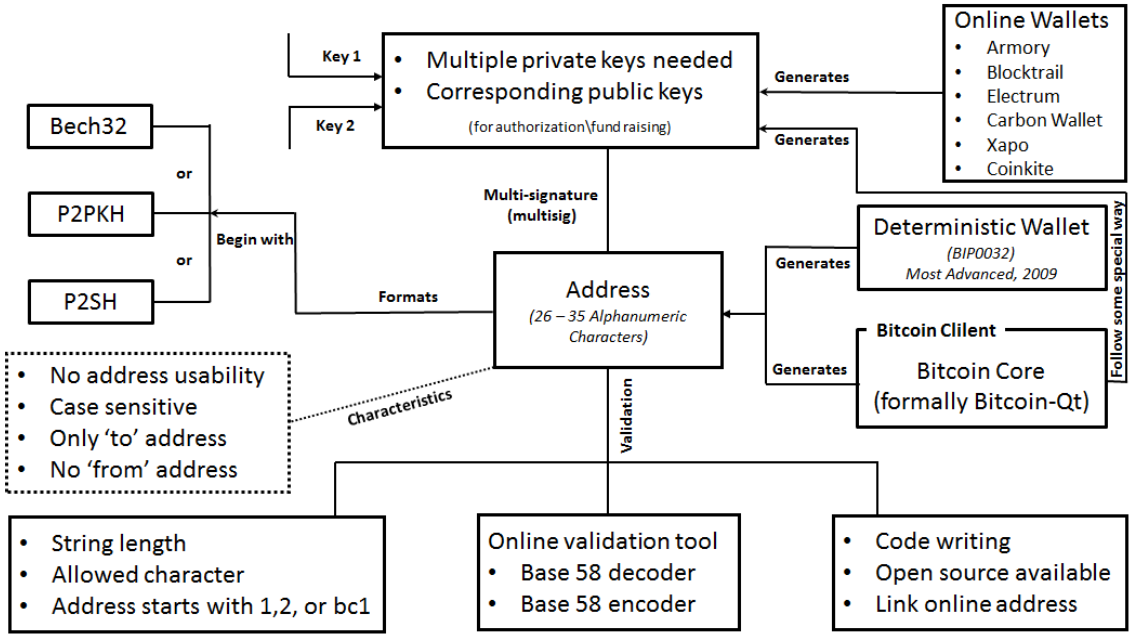


Fig. 2. Cryptocurrency address format, characteristics, generation, and validity.

The final binary address is constructed by appending the checksum, as shown in Eq. (13).

Finally, the binary address is encoded using the Base58 alphabet (A_58) to produce the human-readable address (A_addr).

$$A_{addr} = \text{Encode}\mathcal{A}58(A_{bin}) \tag{14}$$

The human-readable address is generated through Base58 encoding as per Eq. (14).

B. Genesis Block

The Genesis Block is the first block in a blockchain. In the case of the Bitcoin Blockchain, it was initially numbered '1', but modern versions start with '0'. This block is hard-coded and serves as a trusted starting point for the participants. It is an unspendable subsidy that does not refer to a previous block. The genesis block is indexed as 0 in modern implementations.

C. Unspent Transaction Output: UTXO

UTXO is an indivisible chunk that currently represents the "Right to Spend" of someone at a given moment and is recorded on the blockchain. Every received currency chunk is recorded as a UTXO in the blockchain and is scattered throughout the blockchain, even from a particular owner. The Wallet scans the blockchain and calculates the user's balance. The UTXO data on the blockchain is stored in a database table in local memory or in a permanent storage called the UTXO set or UTXO pool.

D. Transaction Format

The UTXO consumed by a transaction (i.e., payment or a transaction fee) is termed transaction input, while transaction output is the UTXO created by a transaction (i.e., change of "Right to Spend" authority or small money returned as change by the spending money with a large valued UTXO).

E. Transaction Fee

A transaction fee is a charge levied on users for each cryptocurrency transaction. This fee is collected to process the transaction on the network. The transaction fee is calculated on the basis of the transaction size in kilobytes. Some wallets automatically calculate and include the transaction fee. Therefore, the user needs to choose a wallet that suits their needs. Additionally, it is crucial to note that the wallet must source a set of UTXOs that add up to sufficient currency (the transaction amount plus the transaction fee). Manual accounting for the transaction fee is recommended when wallet-level fee estimation is unavailable.

F. Merkle Tree

Merkle Root comes up with a Merkle tree when a repeated encryption process forms a tree structure and ends up at a single hashed value. In Ethereum, the state's information is stored in the *Merkle Particia*. It is for preventing the Block from any counterfeit. The *Merkle Path* in the Merkle Tree helps a node find a particular transaction quickly and easily. The conditional state update of the blockchain is represented in Eq. (15):

$$\mathcal{C}_{t+1} = \begin{cases} \text{Append}(\mathcal{C}_t, B_{\text{new}}), & \text{if } \left(\sum_{m_j \in M} V(B_{\text{new}}, m_j) \right) \geq \frac{3}{4}|M|, \\ \mathcal{C}_t, & \text{otherwise.} \end{cases} \tag{15}$$

V. Proposed Solution

In the proposed idea, only some miners participate in the nonce generation and hence do not pick up the transaction from the unconfirmed pool. Each miner generates a random number, and their timer is set to it. The one with the lowest random number value can pick up the transaction. To reduce miners' computational load and save verification time, the relevant data is communicated to the specific node. The selection of this particular node is on the part of the selected miner. The selected miner generates a random number from the range of the total number of nodes and sends the required data to that selected node. The selected node verifies the shared transaction and performs the necessary hashing, culminating in the generation of the Merkle tree that ends at the Merkle root. This selected node shares the data with the head miner. The head miner solves the puzzle for generating the nonce value (explained in the subsection Nonce Generation).

This contrasts with the traditional nonce-value generation competition among all miners, which consumes and wastes computational power and delays block generation. Next to it is the new block generation by the head miner. This newly generated block is appended to the existing blockchain and shared among miners for validation. On having validity confirmation from $\frac{3}{4}th$ of the total miners, this blockchain is replicated to all the miners. A transaction confirmation message is sent to the transaction's direct stakeholders. The detailed workings of the proposed scheme, along with the architectural and communication models, are presented in subsequent subsections. Table II gives some commonly used terms and their descriptions. A schematic diagram of the consensus algorithm-based scalability solution for blockchain is given in Fig. 3. Also, a detailed presentation of the Merkle Tree-Generation algorithm, which is self-descriptive, is given in Algorithm 1. The components of Fig. 3 are presented separately in Fig. 4 and Fig. 5, and their detailed workings are well explained in the proposed solutions. The consensus condition for a new block to be replicated is based on a threshold of valid confirmations, as shown in Eq. (16):

$$N_{confirm} \geq \frac{3}{4}N \tag{16}$$

where, $N_confirm$ is the number of miners confirming validity, and N is the total number of miners. The selection of a miner for nonce generation is based on the lowest random number, as represented in Eq. (17):

$$S = \{i \in M \mid R_i = \min_{j \in M}(R_j)\} \tag{17}$$

where, S is the set of selected miners, M is the set of all miners, and R_i is the random number generated by miner i. To prevent a miner from choosing or modifying R_i after observing competitors' values, the election can use a lightweight commit-and-reveal process. Each eligible miner first broadcasts a commitment $c_i = H(u_i \| pk_i \| H_{prev} \| e)$, where u_i is a locally generated cryptographic random seed, pk_i is the miner identity, H_{prev} is the previous block hash, and e is the current election epoch. After the commitment window closes, miners reveal u_i, and the score is recomputed as $R_i = H(u_i \| pk_i \| H_{prev} \| e)$. Missing or inconsistent reveals are discarded and may trigger slashing. The total consensus time for the proposed system, T_{NCABS}, is a function of the parallel mining time for the head miner, which is a significant improvement over traditional mining methods, as described in Eq. (18):

$$T_{NCABS} < T_{solo} \tag{18}$$

where, T_solo is the time for a traditional solo mining process.

TABLE II. Some Frequently Used Terminologies and their Description.

Term	Description
$V_{(Block)}$	Block validators/miners
$V_{(Tx)}$	Transaction Validators
$Act_{(nodes)}$	Actors
$Power_{(Comp)}$	Computational power
$Wealth_{(Tkn)}$	Token wealth
$Pool_{(U-Tx)}$	Unconfirmed Transactions Pool
$Pool_{(C-Tx)}$	Confirmed Transactions Pool
$Gen_{(Block)}$	Block Generator

A. Architectural Model

The stakeholder nodes are grouped into three major categories namely; Block Validators $[V_{(Block)}]$, Transaction Validators $[V_{(Tx)}]$, and Actors $[Act_{(nodes)}]$. Computational power $[Power_{(Comp)}]$ and Token wealth $[Wealth_{(Tkn)}]$ are two parameters to set the threshold for categorizing the applicant nodes into Block validators and Transaction validators. Actors are transaction nodes that have installed the Blockchain client and wallet applications to interact with the Blockchain network and perform transactions (sending and receiving currency). Transaction validators are nodes with applicants that meet the lower thresholds for $Power_{(Comp)}$ and $Wealth_{(Tkn)}$. The applicant nodes that meet the upper limit criteria for the threshold values, i.e., $Power_{(Comp)}$ and $Wealth_{(Tkn)}$, are categorized as Block validators. These are also the full nodes. Since their work is heavier and more resource-intensive than that of the transaction nodes (Actors), e.g., due to nonce generation. First, the Transaction is entered into the system and placed in the Unconfirmed Transaction Pool $[Pool_{(U-Tx)}]$. Later, once it reaches back from the Transaction Validators $[V_{(Tx)}]$, it is placed by the miner in its own confirmed transactions pool $[Pool_{(C-Tx)}]$ that is explained in Section V-C1a and

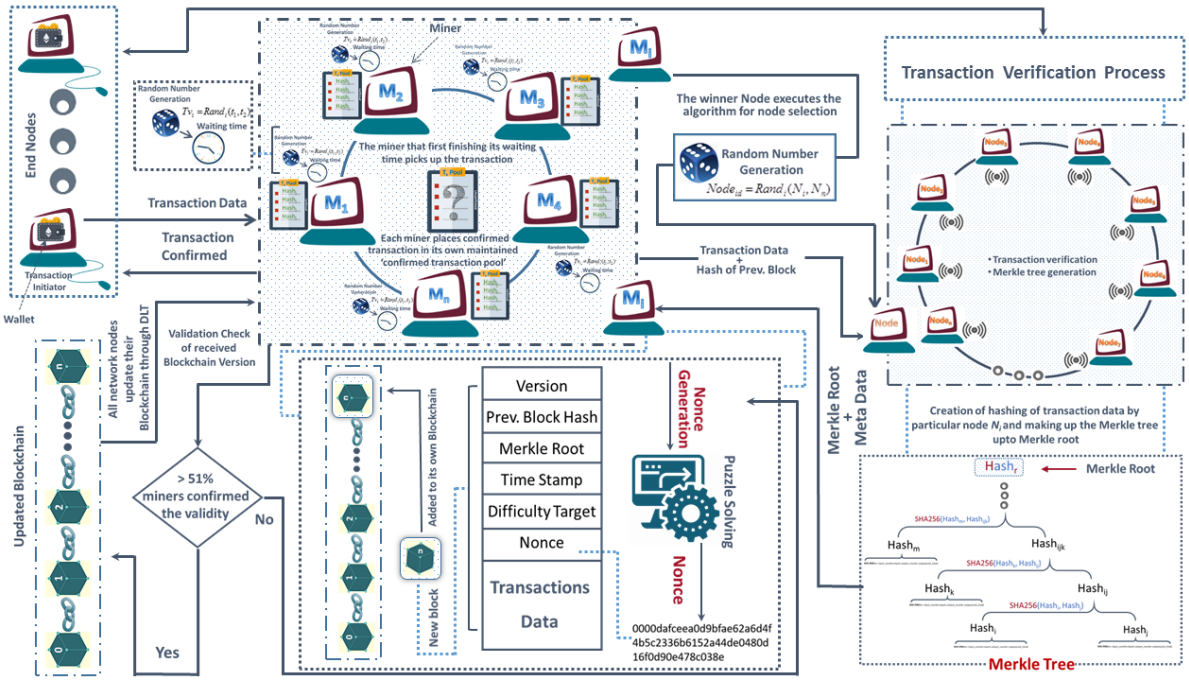


Fig. 3. Consensus algorithm-based scalability solution for blockchain.

are graphically represented in Fig. 4. The description of the terminologies used in this study is tabulated in Table II.

B. Communication Model

Moreover, the higher the number of incoming transactions, the lower the risk of such inconveniences caused by outdated node status. As with each node's response to the transaction verification data, the updated list of alive nodes is shared. So there is very little chance of incorrect information. Every node that sends the transaction verification data back to the head miner will attach the list of live nodes. This list of received messages is then communicated to co-miners. The node collects this information while getting its transactions verified.

C. Detailed Working of the Proposed Solution

The proposed solution operates through a structured consensus algorithm where validation precedes consensus to ensure the running blockchain is the final, agreed-upon version. The process begins when a transaction is placed into an unconfirmed transaction pool, $Pool_(U-Tx)$. A block generator, $Gen_(Block)$, is then selected via a protocol in which all miners generate a random number and set a corresponding timer; the miner with the shortest timer value wins the right to pick up the transaction bunch. This winning miner, or head miner, then delegates the computationally intensive verification task by randomly selecting a dedicated Transaction Validator node, $V_(Tx)$. This selected $V_(Tx)$ node verifies the transaction's legitimacy, syntax, and data structure, and performs the necessary hashing to generate a Merkle tree and its corresponding Merkle root. The validator sends this data back to the head miner, who then solves the puzzle to generate a nonce, creates the new block, and appends it to its local blockchain. This newly updated blockchain is shared with other miners for validation. Upon receiving confirmation from at least 3/4ths of the total miners, it is replicated across the entire network, and a confirmation message is sent to the transaction's stakeholders.

1) Consensus algorithm: The validation precedes the consensus. The consensus protocol ensures that the currently running Blockchain is the final, true version on which at least the system's key stakeholders have agreed. In the proposed consensus algorithm, the Blockchain is replicated to all the miners, once having the validity confirmation from $\frac{3}{4}th$ of total miners. This $\frac{3}{4}th$ ratio is chosen following the renowned previous practices in the literature. Section V-C1c explains it in detail. Before it, nonce generation is the core activity miners perform to make the process fairer, as described in Section V-C1b. After the Block generation and the replication of the Blockchain throughout the network through DLT. The transaction is confirmed and committed. This process is presented in Section V-C1d.

a) Transaction pool: The transaction pushed from the Wallet into the blockchain system is placed in the transaction pool. In Bitcoin, it is named the Mempool. The selected miner picks up the transaction(s) out of the available transactions in the transaction pool [Unconfirmed Transactions Pool $(Pool_{(U-Tx)})$]. Once the transactions are returned from the $V_{(T_x)}$ after confirmation, each miner places them in their in-house maintained pool, the Confirmed Transactions Pool $[Pool_{(C-Tx)}]$. There are some issues of delaying in picking up the transaction from $Pool_{(U-Tx)}$ due to low transaction fees. One solution to it is increasing the transaction fee to push it faster to get it done. SegWit-enabled wallets like Ledger Nano S, Exodus, Segwitaddress, Trezor, or Electrum are worth considering to get faster confirmation from the transaction

Algorithm 1 Merkle Tree Generation

Input: A set of transactions in a block, $T = \{tx_1, tx_2, \ldots, tx_n\}$.
Output: The Merkle Root, M_R.
Definitions:
    Let $H(\cdot)$ be a cryptographic hash function (e.g., SHA-256).
    Let $||$ denote the byte-wise concatenation operator.
    Let L_j be the list of hash nodes at level j of the tree.
function COMPUTEHASH($data$)
    return $H(data)$
end function
function BUILDMERKLETREE(T) ▷ Generate leaf nodes (Level 0)
    Initialize $L_0 \leftarrow \emptyset$
    for $i = 1$ to n **do**
        $l_i \leftarrow$ ComputeHash(tx_i)
        Append l_i to L_0
    end for
    $j \leftarrow 0$ ▷ Iteratively build the tree until one root remains
    while $|L_j| > 1$ **do**
        Initialize $L_{j+1} \leftarrow \emptyset$
        Let $N_j = |L_j|$ ▷ Iterate over pairs of nodes in the current level
        for $k = 1$ to $\lceil N_j/2 \rceil$ **do**
            Let the left child hash be $h_{left} \leftarrow L_j[2k-1]$.
            Let the right child hash be $h_{right} \leftarrow \begin{cases} L_j[2k] & \text{if } 2k \leq N_j \\ L_j[2k-1] & \text{if } 2k > N_j \text{ (Handles odd number of nodes)} \end{cases}$ ▷ Compute the parent hash by hashing the concatenated children's hashes
            The parent hash h_{parent} is computed as:

$$h_{parent} = H(h_{left}||h_{right})$$

            Append h_{parent} to L_{j+1}
        end for
        $j \leftarrow j + 1$
    end while ▷ The final list contains the single Merkle Root
    $M_R \leftarrow L_j[0]$
    return M_R
end function

pool. In the experimental setup, submitted transactions are processed according to their priority, and no external acceleration application is used to push selected transactions ahead of the normal queue. A concise depiction of transaction pools is shown in Fig. 4.

b) Nonce generation: A nonce is a number that results from repeatedly hashing the input until it meets a certain difficulty threshold. In a Block, *nonce* generation is a major task performed by miners and takes most of their time. The process of generating nonces is called mining, and miners mainly do it. It is also called *puzzle solving* because it is a completely random, trial-and-error process that continues until the eligible signature is reached and the required difficulty level is met. It could take many forms, ranging from spaces to question marks to numbers, periods, capital and lowercase

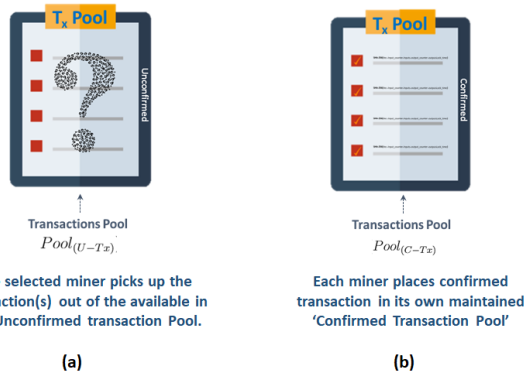


Fig. 4. (a) Unconfirmed transaction pool, (b) Confirmed transaction pool

letters, and digits. Once *nonce* is generated, it is disseminated among other miners along with its key for verification. This also indicates to other miners that they should pause and stop proceeding to actual block generation. Since each block generation process requires the *previous block hash* as a prerequisite, if all nodes start generating blocks, it can lead to confusion about the *previous block hash*. This sharing of the newly generated *nonce* also confirms the miner's message and prevents it from being misleading. This successful node completes the block generation process by setting its *previous block hash* parameter to the latest record in the Blockchain. The locally upgraded Blockchain at the particular miner is sent to the other miners for validation. This received message is a *Go Ahead* call for two actions at the message recipient *i.e.*, Miners. 1) Validation check of the received Block version and respond to the outcome. 2) Proceed to the block generation process.

c) Block generation: The Block Generator $[Gen_{(Block)}]$ is selected by following a special protocol. All the miners generate a random number and set their timers accordingly. The miner with the least timer value takes the right to pick the transaction bunch from the $Pool_{(U-Tx)}$. From a practical perspective, the miner completes its waiting time, which is equal to the randomly generated value, wins the competition, and takes the turn to start the Block generation process, first by picking the transaction(s). A message disseminated to all other miners from this winning miner stops them from waiting further. The message sender miner, *i.e.* the previously winning miner, does not take part in the next competition and will wait for a message from the winner node of the next turn to take part therein. The message also contains its random number value. All the recipient miners wait for this time, then generate a random value as previously presented and discussed on completion of this phase of transaction bunch picking. The transaction verification process is initiated by the $V_{(T_x)}$ node. Once all the required data for a Block is collected and produced, the selected $Gen_{(Block)}$ node proceeds to its generation. Here, another important point needs attention, *i.e.* a serious clash surely will be raised on getting the previous *Block Hash* and including it in the required data for generating the *nonce* and hence the Block generation. Suppose two miners proceed with Block generation using the same *previous hash value*. In

that case, it will disrupt the Blockchain, leading to multiple forks that would be impossible to resolve later. Hence, it is critically important to focus on it and to resolve it.

The fairness of this selection depends on making the random values unpredictable before the election and verifiable after the election. NCABS therefore assumes that random numbers are generated using a cryptographically secure pseudo-random number generator and are bound to the previous block hash and the current epoch through the commit-and-reveal process described above. This binding prevents a miner from reusing a favorable value across epochs, and the public reveal enables other miners to recompute the score before accepting the head-miner result. If two miners obtain the same score, the tie is resolved deterministically by selecting the miner with the smaller hash of $(pk_i \| H_{prev} \| e)$.

The miner that starts the process of generating the *nonce* value up to the Block generation disseminates a message to stop the other miners and to start the same phase in parallel due to the issue of repeated usage of *previous block hash*. During miner selection, it is assumed that candidates must meet a minimum computational power threshold; hence, all miners are capable of generating the Block with proper efficiency. Once the Block is generated and appended to the local blockchain, it is communicated to miners for sufficient validation. On having validity confirmation of $\frac{3}{4}th$ of miners, it is replicated to the blockchain network through DLT. After a successful Block generation, the other miners begin the next Block generation using the *previous block hash*. This is a repetitive process. A concise depiction of Key processes in Block generation by the Block generator and its interaction with the transaction validator module is shown in Figure 5.

d) Transaction verification: Transaction verification is at the end of Transaction Validators' Cluster (TVC), comprising of $(V_{(Tx)})$ nodes. The miner that picks the transactions bunch generates a random number having value from v_i to v_{n-1} (n is the number of nodes in ***TVC***). The $V_{(T_x)}$ corresponds to the outcome number of this random number generation assigned to the transaction bunch. The $V_{(T_x)}$ picks a transaction from the transaction bunch at a time and verifies it by following the number of encapsulated tasks. Firstly, transaction legitimacy is checked, i.e., whether the transactions are legal, not malicious, and not double-spent. In addition, the transaction syntax and data structure, the input and output values, the transaction size (i.e., must be less than the block size), the transaction value (i.e., must be between the minimum and maximum limit values), and the signature limit are verified. When a transaction is marked as correct, the next step is checking that its matching transaction (i.e., child transaction) exists, and if so, then both the parents and child transactions are processed further by placing them in the transaction pool; otherwise, it is put in the *Orphan Transaction Pool*. Once it passes the initial screening, it is hashed. This process is repeated for all transactions in the transaction bunch, and the *Merkle tree* is completed, yielding the *Merkle Root*. To maintain privacy while obtaining the desired data from the neighboring node to verify the payment, *Bloom Filter* is part of the transaction verification process. It comprises various patterns to aggregate data as defined in *BIP37*. Fig. 6 depicts the transaction processing flow as a block diagram, starting with the availability of currency in the wallet and ending with committing the transaction to lock the wallet. This figure presents another schematic diagram (at the abstract level) of Fig. 3.

2) Slashing: The cheater participant (miner and transaction verified) is penalized by having its stake confiscated. This removal of the cheater's stake reduces the token supply on the blockchain, thereby elevating the token price at the individual level. This is another way for bad actors to cover up the impact of their dishonesty. Also, this removed stake is non-transferable and is directly dumped for no use. The full process of leader election, transaction validation, mining, and consensus is detailed in Algorithm 2.

The transaction verification algorithm details the procedure performed by a dedicated validator node. It checks the validity of each transaction in a batch and computes the final Merkle Root, as shown in Algorithm 3.

D. Security Considerations and Adversarial Model

NCABS is designed primarily for a permissioned or consortium blockchain setting in which validator identities are known and eligibility is controlled through the computational-power and token-wealth thresholds described in the architectural model. This membership assumption is important for security. A Sybil adversary cannot freely create unlimited validator identities unless it also satisfies the admission criteria for each identity. The slashing rule further discourages strategic behavior by penalizing miners or transaction validators that submit inconsistent election reveals, invalid validation results, or malformed block data.

The 75% confirmation threshold provides a Byzantine-fault-oriented safety condition. In a validator set of size N, an adversary controlling fewer than one-fourth of the confirming miners cannot independently finalize an invalid block because the block must receive at least $\frac{3}{4}N$ valid confirmations. This threshold also reduces the likelihood of forks because blocks generated from stale or conflicting previous hashes do not receive sufficient confirmations. However, a large colluding majority remains a serious risk in any consensus protocol; 51% attacks have been shown to affect cryptocurrency value and trust [24]. Therefore, NCABS should be interpreted as resistant under the stated permissioned-membership and honest-majority assumptions, rather than as a complete defense against majority capture in an open permissionless network.

Double-spending attacks are addressed at the transaction-verification layer. The selected transaction validator checks transaction legitimacy, syntax, input and output values, signatures, transaction size, and double-spending conditions before generating the Merkle root. A transaction batch that fails these checks is rejected before block assembly. The current work evaluates these mechanisms from a protocol-design and performance perspective; systematic adversarial simulation, including coordinated Sybil attempts, double-spending campaigns, selfish mining, and majority-capture scenarios, remains part of the future validation plan.

E. Complexity Analysis

The computational complexity of NCABS is reduced compared to PoW due to selective miner participation. In

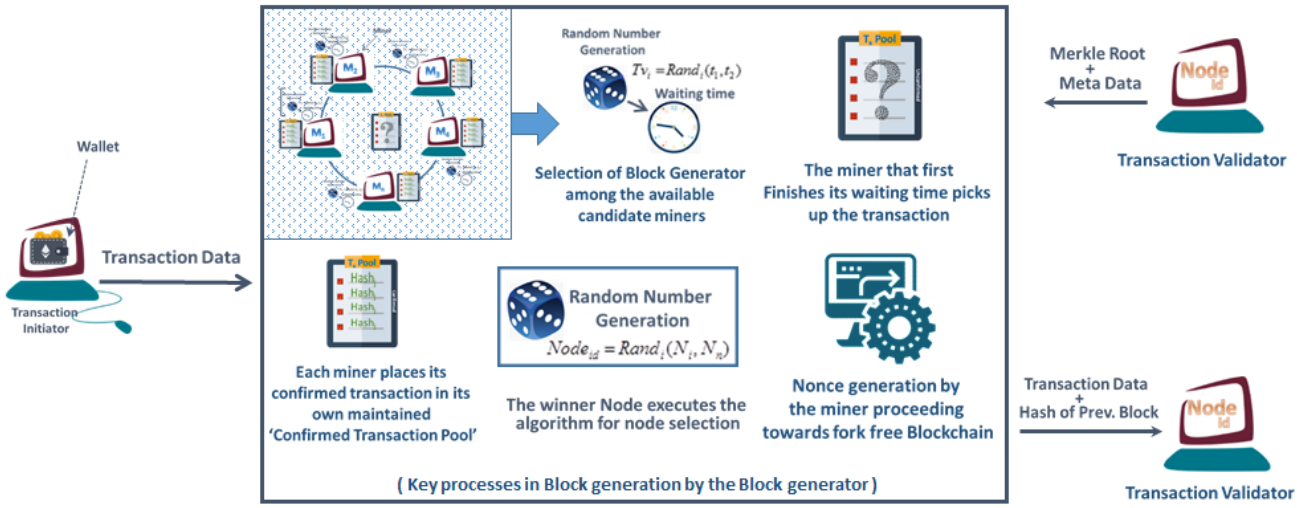


Fig. 5. Key processes in block generation by the Block generator and its interaction with transaction validator module.

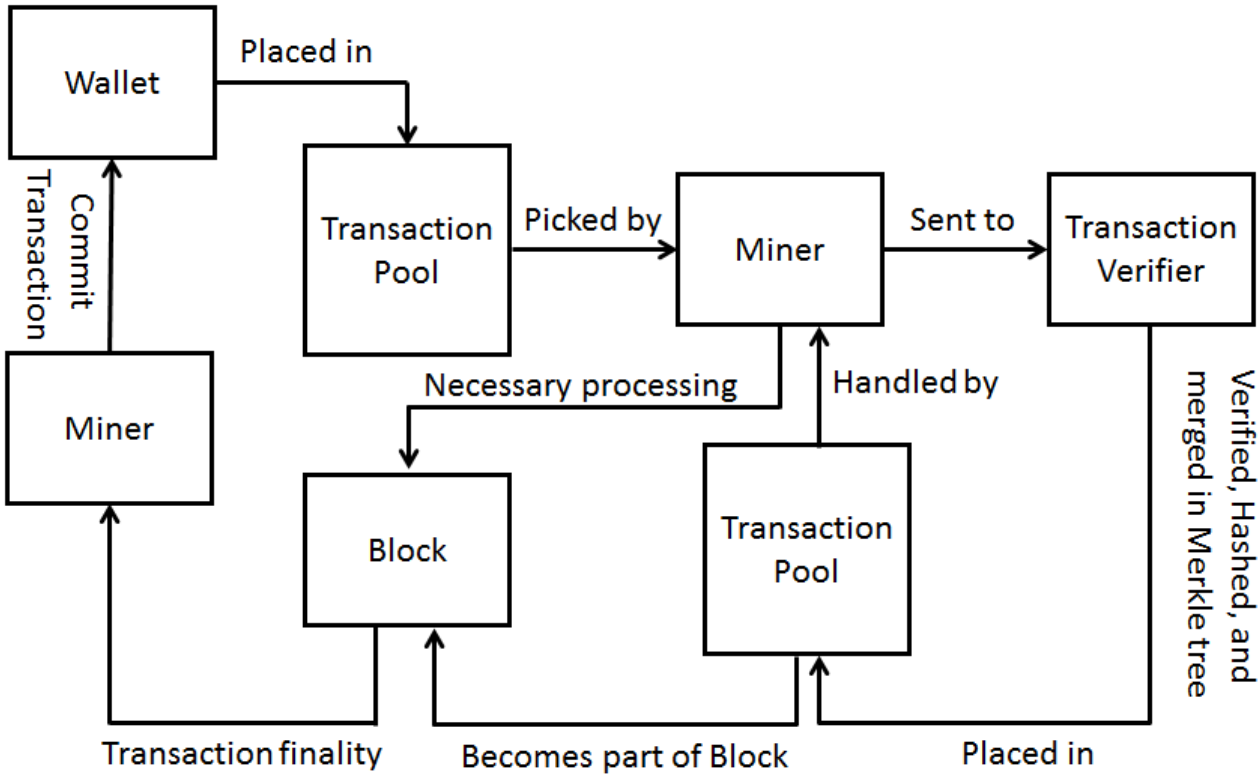


Fig. 6. Transaction processing flow in the proposed blockchain.

traditional PoW, $O(N)$ miners compete to solve the same block-generation problem. In NCABS, only $O(k)$ selected miners participate in the active mining phase, where $k \ll N$. The corresponding computational reduction ratio is therefore approximately k/N. From the communication perspective, PBFT has $O(N^2)$ message complexity because of its all-to-all

Algorithm 2 Proposed Consensus and Block Creation Process

procedure BLOCKCREATIONPROCESS($H_{prev}, Pool_{U-Tx}, \mathcal{M}, \mathcal{V}$)
▷ **Phase 1: Block Generator Election**
  Initialize timers $T_i \leftarrow \infty$ for all $M_i \in \mathcal{M}$
  for each miner $M_i \in \mathcal{M}$ in parallel **do**
    Generate random seed u_i using a CSPRNG
    Broadcast commitment $c_i \leftarrow H(u_i \| pk_i \| H_{prev} \| e)$
    Reveal u_i after the commitment window closes
    Compute score $r_i \leftarrow H(u_i \| pk_i \| H_{prev} \| e)$
    Set timer $T_i \leftarrow r_i$
  end for
  Wait for the first timer to expire
  $M_{gen} \leftarrow \arg\min_{M_i \in \mathcal{M}}(T_i)$
  M_{gen} broadcasts a halt signal to all other miners in $\mathcal{M}$
▷ **Phase 2: Transaction Validation Delegation**
  $\mathcal{T}_{bunch} \leftarrow M_{gen}$ selects a batch of transactions from $Pool_{U-Tx}$
  $V_{Tx} \leftarrow$ randomly select a validator from $\mathcal{V}$
  Send $\mathcal{T}_{bunch}$ to the selected validator V_{Tx}
  $(\mathcal{T}_{valid}, H_{Merkle}) \leftarrow$ VERIFYTRANSACTIONS($\mathcal{T}_{bunch}$)
  if H_{Merkle} is null **then**
    return Block Generation Failed (Validation Error)
  end if
▷ **Phase 3: Mining (Proof-of-Work)**
  Initialize $Nonce \leftarrow 0$
  $H_{header} \leftarrow$ CreateHeader(H_{prev}, H_{Merkle}, timestamp, ...)
  while $H(H_{header}, Nonce) > D_T$ **do**
    $Nonce \leftarrow Nonce + 1$
  end while
▷ **Phase 4: Block Assembly and Consensus**
  $B_{new} \leftarrow$ AssembleBlock($H_{header}, Nonce, \mathcal{T}_{valid}$)
  M_{gen} broadcasts B_{new} to all miners in $\mathcal{M}$
  Initialize $confirm_count \leftarrow 1$ ▷ M_{gen} votes for its own block
  for each miner $M_j \in \mathcal{M} \setminus \{M_{gen}\}$ **do**
    if M_j validates B_{new} as correct **then**
      $confirm_count \leftarrow confirm_count + 1$
    end if
  end for
  if $confirm_count \geq \frac{3}{4}|\mathcal{M}|$ **then**
    Append B_{new} to the global blockchain
    return B_{new}
  else
    Discard B_{new}
    return Block Generation Failed (Consensus Not Reached)
  end if
end procedure

voting structure, whereas NCABS reduces the active consensus communication to approximately $O(kN)$. Hence, NCABS reduces both computational and communication overhead when the selected active set remains much smaller than the complete node set.

Algorithm 3 Transaction Verification and Merkle Root Generation

procedure VERIFYTRANSACTIONS($\mathcal{T}_{bunch}$)
  $\mathcal{T}_{valid} \leftarrow \emptyset$
  for each transaction $Tx_i \in \mathcal{T}_{bunch}$ **do**
    if $\Phi(Tx_i)$ is true **then**
      $\mathcal{T}_{valid} \leftarrow \mathcal{T}_{valid} \cup \{Tx_i\}$
    end if
  end for
  if $\mathcal{T}_{valid}$ is not empty **then**
    $H_{Merkle} \leftarrow$ MerkleRoot($\mathcal{T}_{valid}$)
    return $(\mathcal{T}_{valid}, H_{Merkle})$
  else
    return $(\emptyset, \text{null})$
  end if
end procedure

VI. EXPERIMENTATION AND PERFORMANCE EVALUATION

A. Performance Metrics

The following metrics are used to evaluate the performance of the proposed model [see Eq. (19) to Eq. (21)]:

$$\text{Throughput} = \frac{\text{Number of Transactions}}{\text{Time}} \tag{19}$$

$$\text{Latency} = \text{Time per Block Confirmation} \tag{20}$$

$$\text{Efficiency} = \frac{\text{Throughput}}{\text{Resource Usage}} \tag{21}$$

Each experiment is repeated 10 times, and average values with standard deviation are reported to ensure robustness. To validate the performance of the proposed NCABS consensus mechanism, a series of experiments was conducted and compared against the Practical Byzantine Fault Tolerance (PBFT) algorithm. The evaluation was performed using a setup that included specialized GPU chips, the CGminer software package, and membership in the Poolin online mining pool. The core performance metrics for this comparative analysis were throughput, network latency, network scalability, and responsiveness. Experiments were conducted across a range of network node densities (40–140) to assess the system's behavior under increasing load. The results consistently demonstrated that NCABS offers superior performance, achieving higher throughput, lower resource utilization, a higher transaction commit rate, and lower latency than PBFT. For example, NCABS reduces aggregate communicated messages by approximately 45% in the scalability experiment, and the responsiveness experiment shows a latency reduction from about 310 ms for PBFT to about 195 ms for NCABS at 40 nodes.

B. Theoretical Evaluation

Usually, miners experience latency when hearing about transaction validation at different points in their life cycle. Furthermore, different miners select transactions based on their convenience (e.g., transaction fees) and begin validating them to perform the activities required for block generation. But only one block is accepted based on consensus, to be appended to the blockchain and then replicated across the network of miners using *Distributed Ledger Technology*. Both processes are time-and computational resource-intensive. In the proposed solution, both issues are addressed by assigning the transaction verification role to a dedicated node, *i.e. Transaction verifier*. Therefore, miners do not need to remain engaged in this process, and these resources are better used to keep the system up and running.

The input transactions are readily picked by the miners and put in the processing queue for further assignment to the transaction verifier. This utilizes available resources efficiently and effectively, increasing the system throughput *i.e.*, more transactions are processed and reach the finality state per unit time. Miners have more time to dedicate to block-generation processes, such as nonce generation, without competing with one another, resulting in more efficient resource utilization and higher throughput. A single transaction is picked by only one miner rather than multiple miners (as is practised in other blockchains). It reaches finality with a high probability, as there is no risk or uncertainty of dropping or deselecting the generated block at the end. What comes in the flow will gain the finality status. The verifying nodes are well aligned with their assigned tasks, without redundant effort on the same transaction (which may become part of multi-transaction bunches picked by various miners on their turn). *Transaction verifier* works with one miner at a time, while one miner can assign transaction bunches to more than one transaction verifier. This further contributes to effective resource utilization with better block throughput.

During the block generation, the winning miner prevents the other miners from completing their waiting time to pick up the transaction batch. This algorithm has twofold benefits. It saves time wasted unnecessarily and contributes to better system throughput. Secondly, it helps them autonomously synchronize and enter the transaction-picking phase again almost simultaneously, *i.e.*, a fair deal. By using the *temporal analysis*, the introduced nonce-generation process results in the block generation taking a bit longer. If the set procedure in Bitcoin and other blockchains is followed, carrying out the process simultaneously is more time-consuming. Although the temporal impact is not well appreciated in the proposed mechanism, it is far less than the usual process, as mentioned before. Selecting transactions redundantly across multiple miner picks also delays the block generation process.

C. Implementation Detail

Specialized GPU chips, sufficient cooling for the hardware, an always-on internet connection, a mining software package, and membership in both an online cryptocurrency exchange and an online mining pool are the major implementation parameters. The market is well-supplied with crypto-mining software packages, including CGminer, Bitminter, BFGminer, Multiminer, and EasyMiner. CGminer is used as the representative mining software in this implementation because it supports multi-threaded and multi-pool mining configurations [25]. A group of crypto-miners who combine their computational resources over a network can increase their collective computational power. Poolin, Autopool, Slush Pool, Bitfury, and FZpool are among the widely used mining pools. Poolin is used as the representative mining-pool setting in this study and was reported to mine approximately 20% of blocks at the time considered.

1) Performance metrics: To validate the proposed consensus mechanism's performance, it is compared with PBFT across various evaluation parameters. Those are summarized into throughput, network latency, network scalability, and responsiveness. The results are presented from Fig. 7 to Fig. 17. The detailed discussion is given in subsequent sections.

a) Throughput and network latency: To measure the concurrency in the Blockchain system, an important metric is transaction throughput per Second $T_(TPS)$, i.e., $P = \frac{T_(sum)}{\delta t}$ where $T_(sum)$ is block time and δt is the total transaction count during that block time. It reflects the number of transactions packaged per unit of time. It states that the higher the throughput value, the greater the consensus efficiency and the greater the potential for transaction processing. Hence, $Throughput \propto ConcurrencyEfficiency$. Transaction latency is the time between submitting a transaction and its confirmation. It measures the blockchain's consensus algorithm's computational efficiency, communication efficiency, and runtime. Reduced latency will increase transaction confirmations and decrease the risk of blockchain bifurcation.

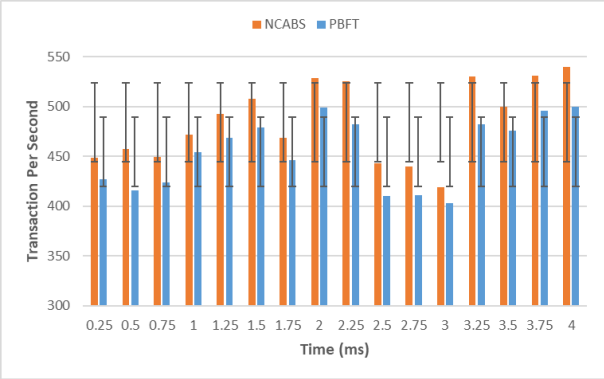


Fig. 7. Average results of throughput - time(ms) vs. transaction per second.

$$Trans_{(Latency)} = Block_{(Conf)}t - Trans_{(Init)}t$$

where, $Trans_{(Latency)}$ is the transaction latency that is the difference between the block confirmation time $Block_{(Conf)}t$ and the transaction initiation time $Trans_{(Init)}t$. Fig. 7 shows the average result of throughput, time vs. transactions per second, and reflects the outperforming of NCABS to PBFT. The variability in results over time indicates changes in the nodes' status. In NCABS, $Power_{(Comp)}$ and $Wealth_{(Tkn)}$ are two parameters used to categorize nodes into Block validators and transaction validators. The status change also affects throughput, as shown in Fig. 7 and Fig. 8. This is also shown in Fig. 9, which compares the change in throughput to the transaction commit rate (Fig. 10). The same also impacts the

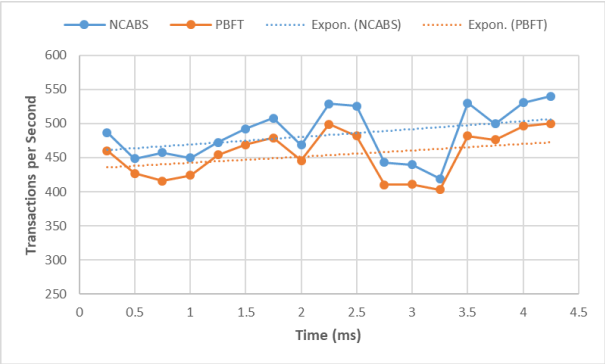


Fig. 8. Exponential trend of competing algorithms on throughput, time(ms) vs. transaction per second.

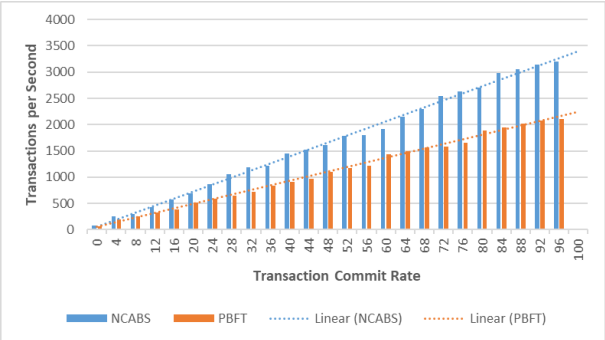


Fig. 9. Transaction commit rate vs. transaction per second for the competing algorithms.

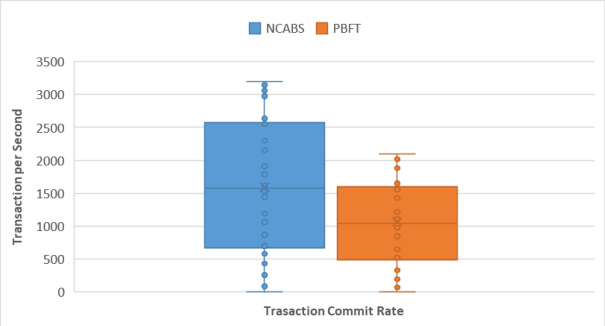


Fig. 10. A summarized view of the throughput comparison of NCABS and PBFT.

graph representing the blockchain network resource utilization, as in Fig. 11. Consequently, this status change also affects message communication since it increases inter-node message passing for node status reallocation. That is reflected in Fig. 11 and Fig. 12.

Across all the above-cited figures, the comparison of NCABS with PBFT for throughput and resource utilization as performance evaluation parameters also shows the network's stability. The higher the frequency of switching the node status, the greater its negative impact on the network. The variation in throughput and resource utilization over time is due to varying node involvement. At t_0, a fraction of nodes serve as validators or supervisor nodes, while at t_1, upon the status

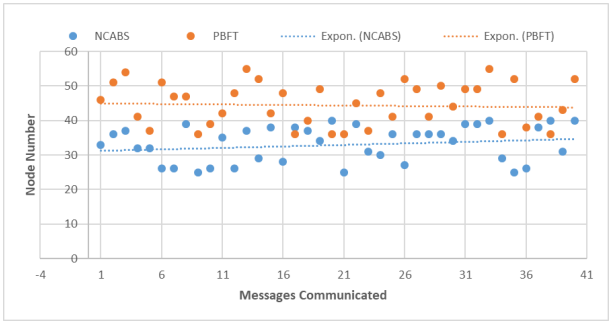


Fig. 11. Blockchain network resource utilization. A comparative view of the per node messages communicated by NCABS and PBFT along with their exponential trend.

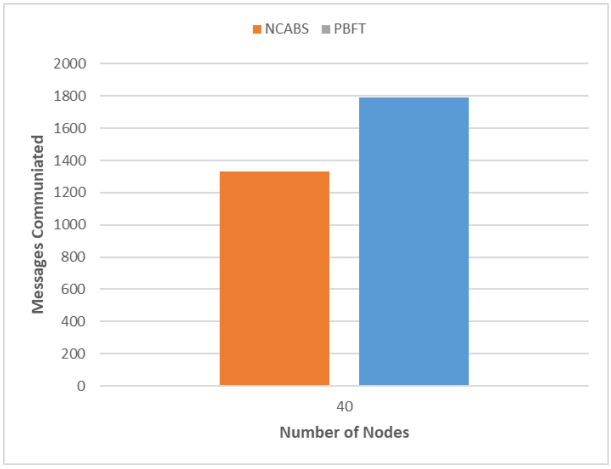


Fig. 12. Blockchain network resource utilization. An aggregated comparative view of the messages communicated in the blockchain network of 40 nodes.

switch, another fraction of nodes assume the same role. This plays a role in the said graph style. In short, as explained above, NCABS achieved higher throughput, lower resource utilization, a higher transaction commit rate, and lower latency than PBFT. Considering Fig. 9, it is clearly evident that the higher the transaction commit rate, the more the throughput, i.e., transactions per second. Over time, throughput increases along with the transaction commit rate for NCABS, and it is higher than that of PBFT. The performance of NCABS is improving compared to PBFT. The reason for this better result is the minimal communication during the validation process. Extending this, the time required to confirm a transaction is also shorter for NCABS than for PBFT.

b) Network scalability: Among many other scalability solutions in Blockchain that require close attention for improvement, throughput and latency are the most important concerns. Adding more nodes significantly increases the time needed to confirm transactions, reducing the number of transactions per second. To address the weakness of the slow transaction verification process, as in the founding systems and their descendant solutions, this job is at the end of the Transaction Validators' Cluster, which comprises selected transaction validator nodes. The same approach is used in PBFT and its variants. One approach to improve the scalability or speed of transaction processing is to design a faster proof-

of-work solution. NCABS uses a method to accelerate the proof-of-work process by parallel rather than solo mining. This decreases the overall burden on the Blockchain network nodes and helps ensure that no more than two miners exert the same effort on a given block.

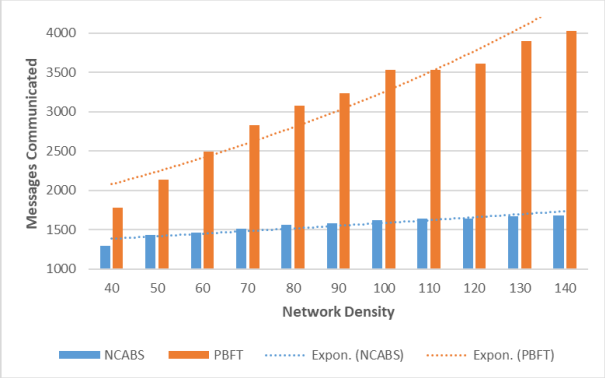


Fig. 13. Comparative results of network scalability over varied network densities in NCABS and PBFT, along with their exponential trend.

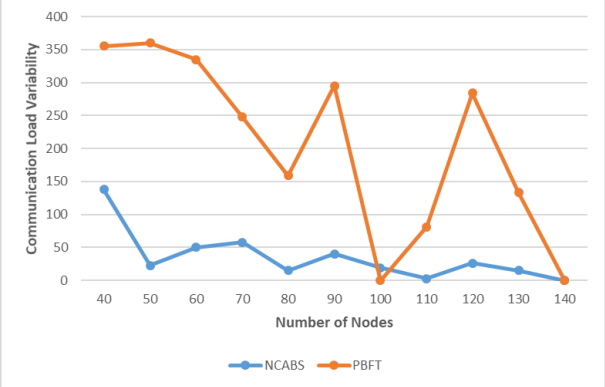


Fig. 14. Communication load variability over varied Number of Nodes in NCABS and PBFT.

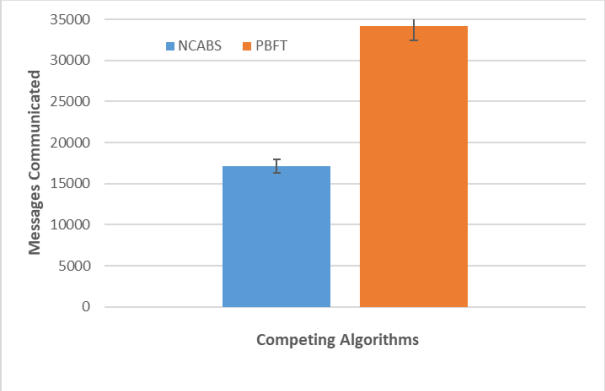


Fig. 15. Network scalability.

Fig. 13, Fig. 14, and Fig. 15 show the network load leading to Network scalability at different network densities. The results shown in the figures are averages across multiple simulations. The results are obtained at varying numbers of nodes, ranging from 40 to 140, with a difference of 10 nodes per experiment. Fig. 15 shows the summary of the inferred and calculated results (averaged) from all the experiments with different blockchain network node densities. It reflects the response of message communication to set up the transaction's confirmation. The total number of messages communicated by the proposed mechanism, NCABS, is approximately 45% fewer than that of PBFT. Referring to Fig. 13, the difference in the number of communicated messages for a network with a node density of 40 and a node density of 140 is huge, especially in the case of PBFT. This difference is due to network instability, specifically to the redundancy introduced by multiple miners mining the same transaction. In contrast, NCABS follows the one-time, one-node policy, as discussed previously, to reduce this communication load. Fig, 13 and Fig. 14 also endorse the network stability scenario. NCABS is far more stable compared to PBFT. Since the variability in the communication load is minimal and gradually improving in various node densities from 40 to 140 in NCABS. In the case of PBFT, this difference increases exponentially. It performs better for small-scale networks but worsens as network scale increases. Also, the exponential trend line for the resulting PBFT graph is much higher and steeper than that for NCABS.

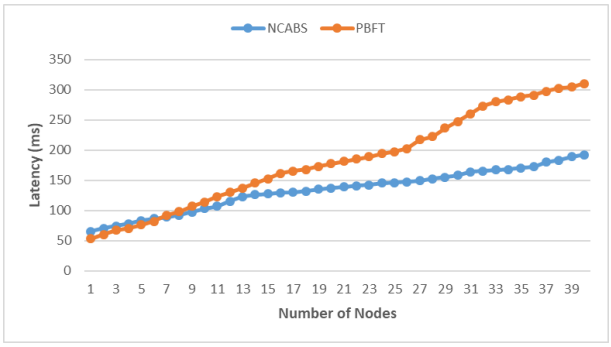


Fig. 16. Network latency under increasing node density for NCABS and PBFT.

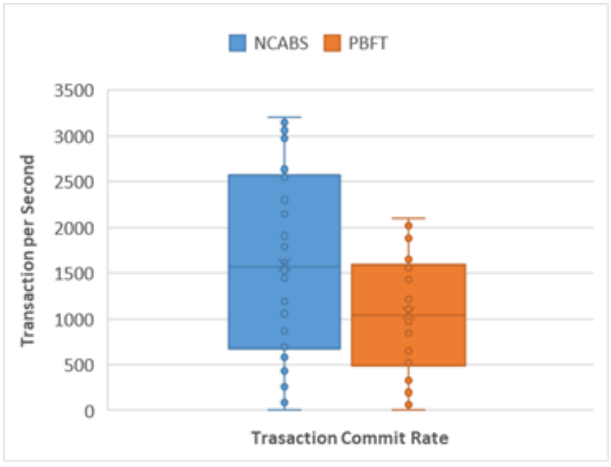


Fig. 17. Aggregated responsiveness comparison using transaction commit rate.

The data in Fig. 13 are also represented as a line graph in Fig. 14, showing far less variability in communication

TABLE III. COMPARISON OF CONSENSUS ALGORITHMS

Algorithm	Communication Complexity	TPS	Groupings of Nodes	Nodes Participation in Consensus Mechanism	Supervision Mechanism	Scalability
PBFT [26]	Average	Low	No	No	No	low
Multi-Layer PBFT [27]	Average	Average	Yes	No	No	Average
NBFT [28]	High	Average	Yes	No	No	Average
GH [29]	High	Low	Yes	No	No	Average
CDBFT [30]	Average	Low	No	Yes	No	Poor
RBFT [31]	High	Average	No	No	Yes	Poor
Algorithm [32]	High	Low	Yes	No	No	Average
PoW [33]	High	Low	No	No	No	Poor
PoS [34]	High	Low	No	No	No	Poor
DPoS [35]	High	Average	No	No	Yes	Average
NCABS (Proposed)	Low	High	Yes	Yes	Yes	Good

load for NCABS compared to that of PBFT. Hence, it results in better network scalability. Since network stability, communication, and network throughput are directly related to network scalability, NCABS delivers far better scalability than its competitor. Fig. 16 and Fig. 17 depict the results of network responsiveness from the perspective of the number of nodes versus latency in milliseconds. In Fig. 16, keeping the node density up to forty, the time interval from transaction submission to transaction confirmation is almost the same for both algorithms for thirteen nodes somewhere in the network. As network latency measures the computational efficiency and runtime of the blockchain's consensus algorithm, throughput is also closely related to it. But later, the latency difference in PBFT increases from node thirteen onward. This decreases its throughput, impacting network scalability. The aggregated responsiveness results are also shown in Fig. 17, indicating that higher latency is associated with lower responsiveness and lower scalability, and vice versa.

D. Comparative Analysis of Typical Consensus with NCABS

Table III compares consensus algorithms, focusing on the extended PBFT ideas. A similar comparison is made by Wu X. *et al.* [36]. The comparison indicators are communication complexity (which also includes communication overhead), Throughput (TPS), nodes' grouping (aka sharding), nodes' participation in the consensus mechanism, supervision of the consensus process, and scalability. The empirical evaluation uses PBFT as the primary baseline because it is the closest directly implemented Byzantine-fault-tolerant comparator in the experimental setup. The remaining consensus protocols are included as qualitative baselines to clarify how NCABS differs from common PBFT variants, PoW, PoS, DPoS, and related group-based protocols. As far as comparing Proof of Work (PoW) and Proof of Stake (PoS), considering the open environment for public blockchain, PoW and PoS are more suitable. Comparing these two, the PoS algorithm prioritizes the highest equity when electing nodes, consumes the least energy, and exhibits a higher degree of centralization than the PoW algorithm, but it lacks security. Its optimized version, DPoS, fine-tunes the number of participating nodes. PoW has stronger decentralisation, a relatively simple implementation, and high fault tolerance, but with the drawbacks of slow transaction speeds, higher energy consumption, and lower throughput [37].

The GH algorithm divides the network nodes into groups, each with a primary node. Local consensus leads to the global consensus achieved by the collective work of only the primary nodes of the designed groups. This increases communication complexity by increasing overhead among interacting nodes. The same is true for NBFT, RBFT, PoW, and its variants, i.e., PoS and DPoS. Hence, it also impacts transaction throughput. Data processing speed has been significantly increased, and security has been improved, but the decentralization aspect needs further optimization. The Paxos, Raft, and PBFT algorithms are better suited to permissioned chains, such as consortium and private chains, that require identity verification and have limited node counts. These consensus algorithms can provide excellent information-processing capabilities, low resource consumption, and tolerance to non-Byzantine errors. However, implementing the Paxos consensus algorithm is more difficult and only suitable for consortium chains and private chains with relatively high security. The Raft algorithm is more practical and performs similarly to Paxos. Still, its shortcomings are similar to those of the Paxos consensus algorithm, and it is only suitable for consortium and private chains with relatively high security. The PBFT consensus algorithm offers strong data processing capabilities and enhanced security and is only applicable to consortium and private chains. In addition, the PoH and CW-PoW consensus algorithms proposed in 2021 are improvements and optimizations of the typical consensus algorithm, and they feature high computational efficiency, low energy consumption, and strong safety and reliability. NCABS is well-suited for the permissioned blockchain for the same reason as PBFT. A detailed experimental-based comparison of these two algorithms is made in the previous section. The comparative analysis results for the algorithms above are shown in Table III.

VII. DISCUSSION

The experimental results show that the main advantage of NCABS comes from reducing redundant work during transaction verification and block generation. In PBFT, consensus communication grows rapidly with the number of participating nodes because voting and message exchange are broadly distributed across the validator set. NCABS narrows the active consensus path by selecting a head miner, delegating transaction verification to a specific validator, and requiring a 75% confirmation threshold before replication. This design

explains why the scalability experiment reports approximately 45% fewer communicated messages than PBFT and why the latency curve remains lower for NCABS as node density increases.

The proposed method is also consistent with the related-work discussion on consensus-based scalability. PoW improves openness and decentralization but wastes computation because many miners compete for a single accepted block. PBFT improves deterministic finality in permissioned environments but can suffer from high communication overhead. PBFT variants such as Multi-Layer PBFT, NBFT, CDBFT, RBFT, and HotStuff-inspired approaches attempt to reduce this overhead through grouping, leadership, or pipelined voting. NCABS follows the same broad goal of reducing redundant coordination, but it does so through controlled parallel mining and explicit separation between transaction validation and block generation. This separation is the main methodological difference from the compared baselines in Table III.

From a deployment perspective, NCABS is best interpreted as a permissioned or consortium-chain mechanism. Its security assumptions depend on controlled validator admission, verifiable random selection, slashing, and an honest-majority condition under the 75% confirmation threshold. These assumptions make NCABS suitable for applications such as supply-chain blockchain systems, where participant identities and computational eligibility can be managed. They also limit direct transfer to open permissionless networks without additional defenses against Sybil identities, selfish mining, coordinated double spending, and majority capture.

The comparison with PBFT provides a direct empirical baseline because PBFT is the closest implemented Byzantine-fault-tolerant comparator in the present evaluation. The qualitative comparison with additional consensus protocols strengthens the positioning of NCABS but does not replace full empirical benchmarking against all modern variants. Therefore, the results should be read as evidence that NCABS improves performance under the stated experimental conditions, while broader adversarial and multi-baseline testing is required before generalizing the claim to all blockchain consensus settings.

VIII. Conclusion

This study presented NCABS, a controlled parallel-mining consensus mechanism for improving blockchain scalability. The conclusions of the study are as follows:

- NCABS reduces redundant mining and transaction-verification effort by separating head-miner selection, validator delegation, block generation, and confirmation into coordinated stages.
- Across ten repeated experiments with node densities from 40 to 140, NCABS achieves better throughput, lower communication overhead, and lower latency than PBFT; in the reported scalability result, aggregate communicated messages are reduced by approximately 45%.
- The comparative analysis indicates that NCABS is particularly suitable for permissioned or consortium blockchain applications where validator identity, computational eligibility, and confirmation thresholds can be managed.

Two limitations remain. First, the current evaluation is simulation/prototype-based and uses PBFT as the direct empirical baseline, while other modern consensus variants are compared qualitatively. Second, the framework assumes minimum computational capability among miners, controlled validator admission, and honest-majority behavior; adversarial experiments involving Sybil behavior, double spending, selfish mining, and majority capture have not yet been conducted.

Future work will extend NCABS through broader empirical benchmarking against additional PBFT variants and modern consensus protocols, real deployment testing, and adversarial evaluation under coordinated attack scenarios.

Acknowledgment

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number: IMSIU-DDRSP2604).

References

[1] R. M. A. Latif, M. Farhan, O. Rizwan, M. Hussain, S. Jabbar, and S. Khalid, "Retail level blockchain transformation for product supply chain using truffle development platform," *Cluster Computing*, vol. 24, pp. 1–16, 2021.

[2] S. Jabbar, H. Lloyd, M. Hammoudeh, B. Adebisi, and U. Raza, "Blockchain-enabled supply chain: analysis, challenges, and future directions," *Multimedia systems*, vol. 27, pp. 787–806, 2021.

[3] M. A. Habib, M. B. Sardar, C. N. F. Faisal, and M. Nasir, "Blockchain-based supply chain for the automation of transaction process: Case study based validation," in *2020 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2020, pp. 1–7.

[4] R. Sujatha, V. Mareeswari, J. M. Chatterjee, A. A. A. Mousa, and A. E. Hassanien, "A bayesian regularized neural network for analyzing bitcoin trends," *IEEE access*, vol. 9, pp. 37 989–38 000, 2021.

[5] M. Abdallah, O. A. Dobre, P.-H. Ho, S. Jabbar, M. J. Khabbaz, and J. J. Rodrigues, "Blockchain-enabled industrial internet of things: Advances, applications, and challenges," *IEEE Internet of Things Magazine*, vol. 3, no. 2, pp. 16–18, 2020.

[6] J. Tian, Y. Li, and C. Jing, "Rvre: a secure and efficient committee-based consensus protocol leveraging reputation power," *The Computer Journal*, p. bxag019, 2026.

[7] L. Yin, Y. Liu, J. Shi, and Y. Zhang, "A general and efficient redactable blockchain scheme based on voting committee," *The Computer Journal*, p. bxag024, 2026.

[8] L. Ke, C. Hsu, M. H. Au, and Z. Xia, "A practical blockchain-based vaccine supply management framework with verifiability and traceability," *The Computer Journal*, p. bxaf115, 2025.

[9] G. A. F. Rebello, G. F. Camilo, L. A. C. de Souza, M. Potop-Butucaru, M. D. de Amorim, M. E. M. Campista, and L. H. M. Costa, "A survey on blockchain scalability: From hardware to layer-two protocols," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2411–2458, 2024.

[10] A. F. Mahdi and F. Rabee, "A blockchain mining proof of work approach based on fog computing virtualization for mobile crowdsensing," in *2024 Third International Conference on Distributed Computing and High Performance Computing (DCHPC)*. IEEE, 2024, pp. 1–9.

[11] G. A. F. Rebello, G. F. Camilo, L. A. C. de Souza, M. Potop-Butucaru, M. D. de Amorim, M. E. M. Campista, and L. H. M. Costa, "A survey on blockchain scalability: From hardware to layer-two protocols," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2411–2458, 2024.

[12] S. Reno and K. Roy, "Navigating the blockchain trilemma: A review of recent advances and emerging solutions in decentralization, security, and scalability optimization." *Computers, Materials & Continua*, vol. 84, no. 2, 2025.

[13] C. K. Lao, S. Zhou, L. Zhang, F. Zhang, and K. Y. Wang, "Unpacking long-latency transactions in ethereum," in *Proceedings of the Workshop on Decentralized Finance and Security*, 2024, pp. 11–20.

[14] S. Lee and H. Kim, "On the robustness of lightning network in bitcoin," *Pervasive and Mobile Computing*, vol. 61, p. 101108, 2020.

[15] X. Wang, H. Zhu, Z. Ning, L. Guo, and Y. Zhang, "Blockchain intelligence for internet of vehicles: Challenges and solutions," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2325–2355, 2023.

[16] L. T. Thibault, T. Sarry, and A. S. Hafid, "Blockchain scaling using rollups: A comprehensive survey," *IEEE Access*, vol. 10, pp. 93 039–93 054, 2022.

[17] J. W. Heo, G. S. Ramachandran, A. Dorri, and R. Jurdak, "Blockchain data storage optimisations: a comprehensive survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–27, 2024.

[18] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 1–24, 2021.

[19] R. Neiheiser, M. Matos, and L. Rodrigues, "Kauri: Scalable bft consensus with pipelined tree-based dissemination and aggregation," in *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, 2021, pp. 35–48.

[20] G. A. F. Rebello, G. F. Camilo, L. A. C. de Souza, M. Potop-Butucaru, M. D. de Amorim, M. E. M. Campista, and L. H. M. Costa, "A survey on blockchain scalability: From hardware to layer-two protocols," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2411–2458, 2024.

[21] B. Lashkari and P. Musilek, "A comprehensive review of blockchain consensus mechanisms," *IEEE access*, vol. 9, pp. 43 620–43 652, 2021.

[22] Y. Shahsavari, K. Zhang, and C. Talhi, "Performance modeling and analysis of hotstuff for blockchain consensus," in *2022 Fourth International Conference on Blockchain Computing and Applications (BCCA)*. IEEE, 2022, pp. 135–142.

[23] I. M. Coelho, V. N. Coelho, R. P. Araujo, W. Yong Qiang, and B. D. Rhodes, "Challenges of pbft-inspired consensus for blockchain and enhancements over neo dbft," *Future Internet*, vol. 12, no. 8, p. 129, 2020.

[24] S. Shanaev, A. Shuraeva, M. Vasenin, and M. Kuznetsov, "Cryptocurrency value and 51% attacks: evidence from event studies," *The Journal of Alternative Investments*, vol. 22, no. 3, pp. 65–77, 2019.

[25] C. Kolivas, "multi-threaded multi-pool fpga and asic miner for bitcoin," Last accessed 12 September 2025. [Online]. Available: https://github.com/ckolivas/cgminer

[26] H. Xu, J. Wu, Q. Pan, X. Guan, and M. Guizani, "A survey on digital twin for industrial internet of things: Applications, technologies and tools," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2569–2598, 2023.

[27] W. Li, C. Feng, L. Zhang, H. Xu, B. Cao, and M. A. Imran, "A scalable multi-layer pbft consensus for blockchain," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 5, pp. 1146–1160, 2020.

[28] J. Yang, Z. Jia, R. Su, X. Wu, and J. Qin, "Improved fault-tolerant consensus based on the pbft algorithm," *IEEE Access*, vol. 10, pp. 30 274–30 283, 2022.

[29] R. Jabbar, E. Dhib, A. B. Said, M. Krichen, N. Fetais, E. Zaidan, and K. Barkaoui, "Blockchain technology for intelligent transportation systems: A systematic literature review," *IEEE Access*, vol. 10, pp. 20 995–21 031, 2022.

[30] W. Issa, N. Moustafa, B. Turnbull, N. Sohrabi, and Z. Tari, "Blockchain-based federated learning for securing internet of things: A comprehensive survey," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–43, 2023.

[31] J. Xiao, T. Luo, C. Li, J. Zhou, and Z. Li, "Ce-pbft: A high availability consensus algorithm for large-scale consortium blockchain," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 2, p. 101957, 2024.

[32] R. Chen, L. Wang, C. Peng, and R. Zhu, "An effective sharding consensus algorithm for blockchain systems," *Electronics*, vol. 11, no. 16, p. 2597, 2022.

[33] W. Issa, N. Moustafa, B. Turnbull, N. Sohrabi, and Z. Tari, "Blockchain-based federated learning for securing internet of things: A comprehensive survey," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–43, 2023.

[34] J. Xu, C. Wang, and X. Jia, "A survey of blockchain consensus protocols," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–35, 2023.

[35] A. Gangwal, H. R. Gangavalli, and A. Thirupathi, "A survey of layer-two blockchain protocols," *Journal of Network and Computer Applications*, vol. 209, p. 103539, 2023.

[36] X. Wu, W. Jiang, M. Song, Z. Jia, and J. Qin, "An efficient sharding consensus algorithm for consortium chains," *Scientific Reports*, vol. 13, no. 1, p. 20, 2023.

[37] M. Wendl, M. H. Doan, and R. Sassen, "The environmental impact of cryptocurrencies using proof of work and proof of stake consensus algorithms: A systematic review," *Journal of Environmental Management*, vol. 326, p. 116530, 2023.

Explainable Neural Network Prediction of Post-COVID-19 Depression via Monte Carlo Simulation

Siham AKIL, Sara SEKKATE, Abdellah ADIB
Laboratory of Intelligence Machine (LIM)
Faculty of Sciences and Technologies, BP 146-28806 Mohammedia, Morocco

***Abstract*—The COVID-19 pandemic has driven a substantial rise in depression, notably in Argentina's post-quarantine period, motivating the need for predictive tools to support timely mental health interventions. This study uses a Feedforward Neural Network (FNN) and Monte Carlo simulations to predict depression scores from key socio-economic and psychological variables—anxiety state, economic income, and education—benchmarked against SVR, GRU, Linear Regression, Decision Tree, and Random Forest. The FNN achieved the best overall performance ($MAE = 4.72$, $RMSE = 6.32$, $R^2 = 0.64$; cross-validated $R^2 = 0.593 \pm 0.048$), while Linear Regression attained the highest R^2 (0.693), suggesting partly linear relationships among predictors. Monte Carlo simulations showed that higher anxiety increased predicted depression, while higher income and education reduced it, underscoring the value of targeted anxiety-reduction and economic-support interventions in post-pandemic mental health policy.**

***Keywords*—Depression prediction; Feedforward Neural Network; Monte Carlo simulations; explainable artificial intelligence; post-COVID-19 mental health; socio-economic factors; SHAP analysis**

I. Introduction

The COVID-19 pandemic has significantly changed the dynamics of global public health, leading to an unprecedented rise in mental health issues [1]. Among these, depression has emerged as a pressing concern, particularly in the extended lockdowns and quarantine measures [2]. Countries around the world have reported high rates of depression, and Argentina, in particular, experienced a significant increase in mental health problems in the post-quarantine period [3]. The psychological consequences of prolonged social isolation, combined with economic instability, has contributed to a serious mental health crisis [4]. This context underscores the need for effective, data-driven approaches to understanding and predicting depression levels, particularly as public health systems continue to be challenged by the long-term consequences of the pandemic [5].

The socio-economic fallout from the pandemic has led to an increase in mental health issues, with depression being one of the most widespread and disabling disorders [6]. The country's pre-existing economic challenges, such as high inflation and unemployment, were further aggravated by the pandemic, pushing many individuals and families into financial distress [7]. These economic stress factors, coupled with the psychological pressure of the pandemic, have led to an increase in cases of depression [8]. The mental health impact of COVID-19 is not limited to those directly affected by the virus; the broader population has experienced a deterioration in psychological well-being due to factors such as economic insecurity, job loss, disruption of educational systems, and social isolation [9]. These interrelated factors make it imperative to explore the socio-economic and psychological determinants of depression in the post-pandemic period.

Previous research has established that mental health is influenced by a range of factors, including economic status, education, social support, and psychological conditions such as anxiety [10]. For instance, individuals with lower economic income are more likely to experience depression, as financial insecurity can aggravate feelings of stress and powerlessness [8]. Similarly, education plays a critical role in mental health, with lower educational attainment often linked to poorer mental health outcomes [11]. Anxiety, a psychological response to uncertainty and stress, has been shown to amplify the symptoms of depression. During the COVID-19 pandemic [12], heightened anxiety levels, driven by fears of infection, job insecurity, and uncertainty about the future, have significantly impacted mental health worldwide. The intersection of these socio-economic and psychological factors presents a complex landscape for predicting depression levels, particularly in a country like Argentina, where the pandemic has reinforced existing vulnerabilities [6].

Accurate prediction of depression in such a multifaceted environment requires sophisticated analytical tools that can capture the interaction between these variables [13]. Traditional mental health assessments, while valuable, often fall short in accounting for the dynamic and interdependent nature of socio-economic and psychological factors [14]. Machine learning (ML), with its ability to process large datasets and uncover hidden patterns [15], offers a powerful solution to this challenge. By leveraging ML techniques, researchers can build predictive models that not only consider individual variables like economic income and education but also capture the complex interactions between these factors and psychological conditions such as anxiety [11].

In recent years, ML models have shown great promise in predicting mental health outcomes, particularly in situations where traditional statistical methods may be less effective [16]. These models are capable of analyzing vast amounts of data, identifying non-linear relationships, and making predictions that can help guide public health interventions [17]. When combined with Monte Carlo simulations, ML models can

further enhance the predictive accuracy by simulating various "what-if" scenarios [18]. Monte Carlo simulations allow researchers to explore the effects of different socio-economic and psychological conditions on depression levels, providing valuable insights into how changes in these variables might influence mental health outcomes [19].

This study aims to leverage these advanced analytical tools, ML and Monte Carlo simulations, to predict depression levels in Argentina's post-quarantine period. Specifically, the study will examine the role of key variables such as anxiety levels, economic income, and education in shaping depression outcomes. By varying these inputs through Monte Carlo simulations, the study aims to understand how different scenarios, such as changes in economic conditions or anxiety levels—impact depression predictions. The research addresses the following research questions:

- How do socio-economic factors, particularly economic income and education, influence depression levels in Argentina's post-quarantine period?
- What is the relationship between anxiety levels and depression, and how can this psychological factor be effectively taken into predictive models?
- How can ML models, combined with Monte Carlo simulations, enhance the prediction of depression in the context of the socio-economic and psychological challenges posed by the COVID-19 pandemic?

By answering these questions, this research provides an important contribution to the field of mental health prediction, particularly in the context of global crises such as the COVID-19 pandemic. The findings from this study are expected to offer valuable insights for policy-makers and healthcare providers in Argentina and other countries facing similar challenges, enabling more targeted mental health interventions. Moreover, the use of FNN model and Monte Carlo simulations represents an innovative approach that can be applied in future mental health research, particularly in complex socio-economic contexts.

The choice of a FNN was motivated by its proven capacity to model complex, non-linear relationships between input features and continuous outcomes, making it particularly well-suited for psychological and behavioral prediction tasks such as depression score estimation. Unlike traditional linear models, FNNs can learn flexible mappings from high-dimensional feature spaces without relying on pre-specified interaction terms or transformation assumptions. Furthermore, the relatively low dimensionality of the input space and the structured nature of the variables made the FNN a computationally efficient yet expressive choice for this regression task.

While the use of ML models for mental health prediction is not new, the novelty of this work lies in three complementary contributions. First, unlike prior studies that typically treat depression prediction as a static regression problem, this study integrates a FNN-based predictive model with Monte Carlo simulation to explicitly quantify how socio-economic and psychological factors—anxiety state, economic income, and education—propagate into depression outcomes under simulated "what-if" scenarios. Second, the model's predictions are made interpretable through SHAP-based explainability, addressing the common trade-off between predictive accuracy and transparency in mental-health applications, where clinical and policy adoption depends on understanding why a prediction was made. Third, this combination of FNN, Monte Carlo simulation, and SHAP explainability is applied, to the best of our knowledge, for the first time to the post-quarantine Argentine context, providing a population- and period-specific predictive tool rather than a generic depression model.

The remainder of this study is structured as follows: Section II discusses relevant studies on mental health, ML models, and Monte Carlo simulations in healthcare. Section III covers the dataset, preprocessing, model architecture, and simulations. Section IV present findings on the impact of key factors like anxiety and income on depression. Section V interprets these results and compares them with existing studies. while Section VI summarizes key insights and future research directions.

II. Literature Review

The global impact of the COVID-19 pandemic has extended far beyond physical health, with mental health becoming a critical concern [1]. The widespread implementation of quarantine measures, social isolation, and economic hardship have contributed to an unprecedented increase in mental health disorders, particularly depression and anxiety [20]. This section reviews the literature on the mental health effects of COVID-19 globally, with a specific focus on Argentina and similar regions, and examines existing models for depression prediction using ML [21]. Furthermore, it discusses the relevance of Monte Carlo simulations in healthcare, particularly for scenario-based analyses.

The mental health consequences of the COVID-19 pandemic have been substantial, affecting populations worldwide [22]. Numerous studies have documented increases in the prevalence of depression, anxiety, and other psychological disorders during and after the quarantine period [23]. Globally, individuals have experienced heightened levels of stress due to health fears, job insecurity, and disruption of social support networks [24]. A study by Lei et al. [25] highlights that depression and anxiety levels increased significantly during the pandemic, with rates of depression among the general population ranging from 6.7% to 14.6%.

In Argentina, the economic and social challenges posed by the pandemic were exacerbated by existing economic instability. Argentina has faced long-standing economic challenges, including high inflation and unemployment, which were worsened by the pandemic's impact. Several studies in Latin America and Argentina have highlighted the direct link between socio-economic hardship and mental health deterioration during COVID-19 [26]. Torrente et al. [27] specifically explored the mental health effects in Argentina, noting a significant rise in depressive symptoms during the lockdown period, with younger adults and those with lower economic income being the most affected.

In addition, the pandemic intensified psychological stressors, such as uncertainty about the future, fear of illness, and social isolation [12]. Badellino et al. [23] found that anxiety and depression rates in Argentina were significantly correlated with the duration of quarantine and the associated social and economic difficulties. The study also identified that individuals

who experienced income loss or unemployment were more likely to suffer from mental health issues. These findings align with global trends, indicating that the socio-economic consequences of the pandemic are key determinants of mental health outcomes [26].

With the increasing availability of mental health data, ML has emerged as a valuable tool for predicting mental health outcomes, including depression [28]. Traditional statistical methods, while useful, often struggle to capture the complex and non-linear relationships between socio-economic, psychological, and demographic factors [29] that contribute to mental health disorders. ML models, however, excel in this area, allowing for the analysis of large datasets with multiple variables and uncovering patterns that may not be immediately obvious.

Several studies have demonstrated the effectiveness of ML models in predicting depression [30]. For instance, Razavi et al. [31] applied ML algorithms to predict depression based on mobile phone usage and behavioral data, achieving high accuracy. Bakkeli et al. [30] has utilized a combination of psychological factors and socio-economic data to build depression prediction models. Sun et al. [32] employed a ML approach to predict depression onset in individuals based on a wide range of clinical and socio-demographic factors, including economic status and education level.

In the context of the COVID-19 pandemic, ML models have been used to predict the mental health impact of various stressors, such as isolation, income loss, and increased anxiety [33]. Hueniken et al. [34] developed a ML model to predict depression based on anxiety and socio-economic factors during the pandemic, showing that these factors were highly predictive of depression risk. These models offer valuable insights into how depression evolves in response to external conditions, making them particularly useful in the context of a global crisis like COVID-19 [35].

Monte Carlo simulations are widely used in healthcare to model uncertainty and assess the potential outcomes of different scenarios [36]. In mental health research, Monte Carlo methods allow for the exploration of how various socio-economic and psychological factors interact to influence mental health outcomes, such as depression [37]. By simulating a range of possible scenarios, researchers can estimate the probability of different outcomes and gain a deeper understanding of the factors driving mental health disorders [38].

In the context of depression prediction, Monte Carlo simulations are particularly useful for exploring "what-if" scenarios [39]. Monte Carlo simulations have also been employed in scenario-based analyses to evaluate the effectiveness of mental health interventions [40].

In this study, Monte Carlo simulations are used in conjunction with ML models to simulate the impact of changes in socio-economic and psychological factors—such as anxiety state, economic income, and education—on depression levels in Argentina during the post-quarantine period. This approach allows for a more comprehensive understanding of how different variables contribute to depression and enables the identification of potential interventions that could alleviate the mental health burden.

III. Methodology

In this section, we detail the methodology employed in this study, which includes data preprocessing, model training, and Monte Carlo simulations. It describe how socio-economic and psychological variables were processed, the steps taken to train a FNN model, and the approach used to simulate various scenarios. The flowchart in Fig. 1 illustrates the entire workflow, from data preparation to the prediction of depression outcomes.

A. Data Description

The dataset titled "Mental Health of People in Argentina Post-Quarantine COVID-19" used in this study was collected by López Steinmetz from 1100 individuals residing in different provinces of Argentina during the post-quarantine period following the COVID-19 pandemic [41]. This period is characterized by significant socio-economic changes and heightened mental health concerns. The dataset includes key demographic, economic, and psychological variables that provide insights into the mental health status of individuals. Table I provides a summary of the variables included in the dataset.

These features were chosen for their known relationship with mental health, and their interaction offers a complex view of depression in the post-quarantine context.

B. Data Cleaning

1) Handling missing values: Before applying ML models, the dataset was cleaned to handle missing values. For continuous variables such as AGE, DEPRESSION, ANXIETY STATE, and ANXIETY TRAIT, mean imputation was selected over alternative strategies (e.g., median imputation or model-based imputation such as k-NN or MICE) because the proportion of missing values in the continuous variables was low (under 5%) and a preliminary inspection of their distributions did not reveal strong skewness or extreme outliers that would bias the mean. Under these conditions, mean imputation preserves the original sample size and the overall distributional shape of each variable while avoiding the additional computational complexity and risk of introducing artificial correlations associated with more elaborate imputation methods. Let $X = x_1, x_2, ..., x_n$ represent the dataset where x_i denotes the features for each respondent. Missing values in continuous variables were replaced by the mean of the available data:

$$x_i{}^{(missing)} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1}$$

where, n is the total number of respondents or data points. $x_i{}^{(missing)}$ is the value used to replace the missing data for feature x_i.

For categorical variables, missing values were replaced by the mode, defined as:

$$x_i^{(missing)} = \underset{v \in \text{values}(x)}{\arg\max}\ \#(v) \tag{2}$$

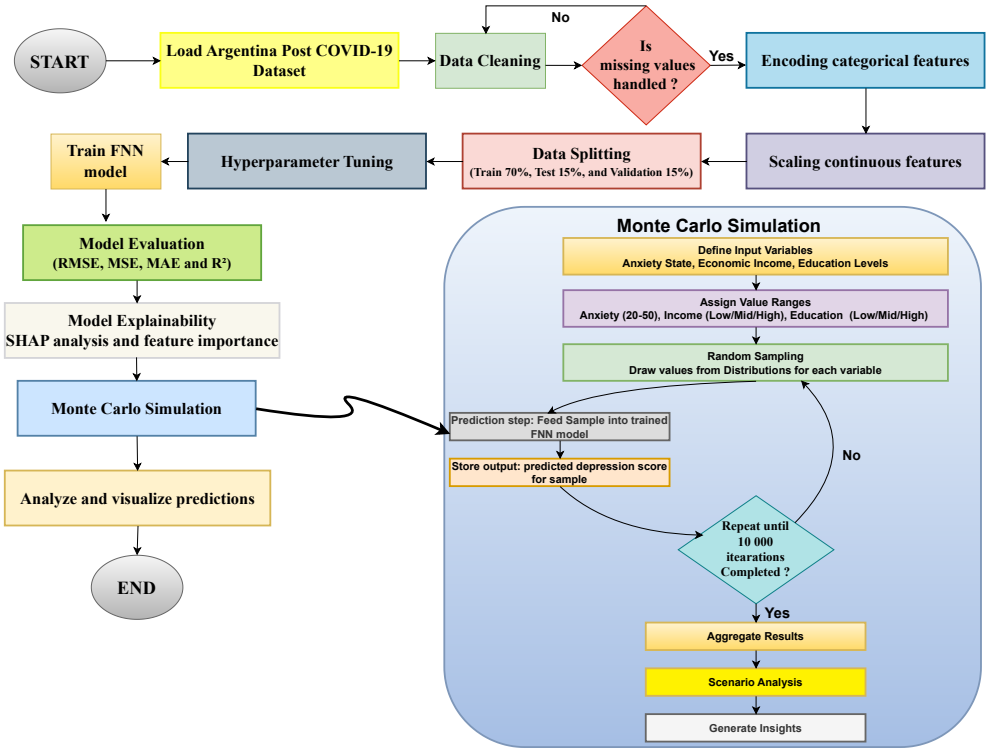


Fig. 1. Flowchart of the proposed methodology.

TABLE I. DESCRIPTION AND DATA TYPES OF VARIABLES

Variable	Description	Data Type
SUB PERIODS	Survey collection phase indicator	Categorical
EDUCATION	Maximum level of education attained by the individual	Categorical
PROVINCE	Province where the individual resides	Categorical
SEX	Gender of the individual	Categorical
AGE	Age of the individual	Numerical
MENTAL DISORDER HISTORY	History of mental disorder diagnosis	Categorical
SUIC ATTEMPT HISTORY	History of suicide attempts	Categorical
LIVING WITH SOMEBODY	Indicates whether the individual lives with someone	Categorical
ECONOMIC INCOME	Whether the individual has economic income	Categorical
DEPRESSION	Depression score (quantitative)	Numerical
SUIC RISK	Suicide risk score	Numerical
ANXIETY STATE	State anxiety score	Numerical
ANXIETY TRAIT	Trait anxiety score	Numerical
REGION	Name of the region where the individual resides	Categorical

where, $\arg\max \#(v)$ refers to the value v that occurs most frequently (the mode) among the available values of the feature x. values(x) is the set of all observed values of the feature x. and $\#(v)$ denotes the frequency or number of times the value v appears in the dataset.

2) Categorical encoding: Categorical variables like SEX, ECONOMIC INCOME, EDUCATION, and MENTAL DISORDER HISTORY were encoded numerically. One-hot encoding was used for multi-class variables, while binary variables were mapped as 0 and 1. For example, the education levels were mapped as:

$$\text{EDUCATION} = \begin{cases} 30 & \text{Completed post graduate} \\ 40 & \text{Incomplete post graduate} \\ 50 & \text{Completed tertiary or university} \\ 60 & \text{Incomplete tertiary or university} \\ 70 & \text{Completed high school} \\ 80 & \text{Incomplete high school} \\ 90 & \text{Completed elementary school} \\ 100 & \text{Incomplete elementary school} \end{cases}$$

One-hot encoding transforms a categorical variable x with k categories into a binary vector $y \in \{0,1\}^k$:

$$y_i = \begin{cases} 1 & \text{if the } i\text{-th category is present} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

3) Data scaling: Continuous variables were scaled using the MinMaxScaler to ensure all variables fell within a consistent range. Let x_i be a continuous variable, it was normalized as follows:

$$x_i^{scaled} = \frac{x_i - min(x)}{max(x) - min(x)} \quad (4)$$

This ensures that the features lie in the range $[0, 1]$, improving the performance of ML models [42]. MinMax scaling was preferred over standardization (z-score normalization) because several of the FNN's input features, after one-hot and ordinal encoding, do not follow a Gaussian distribution and have bounded, interpretable ranges (e.g., anxiety scores, encoded education levels). Rescaling all features to a common $[0, 1]$ range avoids implicitly assigning greater weight to variables with larger numeric ranges and is particularly suited to neural network training, where bounded inputs help stabilize gradient-based optimization and accelerate convergence.

4) Feature engineering: Feature engineering was a critical step in this study, as it allowed for the transformation of raw data into meaningful inputs for both the ML model and Monte Carlo simulations. The key features engineered include:

1) Anxiety state: This feature represents the temporary anxiety levels of respondents. To explore its impact on depression, anxiety state was varied from low (20) to high (50) in increments of 5 during the simulations. This variation allowed the model to predict how increasing anxiety levels would affect depression scores.

$$AnxietyState = [20, 25, 30, 35, 40, 45, 50]$$

2) Economic income: Respondents' economic income was categorized into low, middle, and high income. These categories were used in the Monte Carlo simulations to explore how changes in economic status affect depression predictions. The three categories were encoded as follows:

$$Economic\,Income = \{Low\,Income, Middle$$

$$Income, High\,Income\}$$

3) Education: Education levels were categorized into low (incomplete high school), middle (completed high school), and high (completed university or postgraduate). These categories were used to simulate how different levels of educational attainment influence depression outcomes.

$$Education = \{Low, Middle, High\}$$

By varying these features in different simulations, this allowed for the exploration of explore the role of socio-economic and psychological factors in predicting depression. Interaction terms were also considered during feature engineering, such as the combination of economic income and education, to capture how these factors together affect depression outcomes.

C. Data Splitting

The dataset was split into training (70%), validation (15%), and testing (15%) sets. The training set was used to fit the model, the validation set was used for tuning hyperparameters, and the testing set was used to assess model performance on unseen data. A dedicated validation set was retained, in addition to the 5-fold cross-validation described in Section III G, to support the hyperparameter search described in Section III F. While cross-validation is used to report robust, averaged generalization performance across the full dataset, an independent validation split was needed during the random search procedure to select the optimal configuration efficiently without repeatedly re-running k-fold cross-validation for every candidate combination, which would have been computationally prohibitive given the number of hyperparameter combinations explored. The final model, trained with the selected configuration, was then evaluated on the untouched test set and separately validated through 5-fold cross-validation to confirm that its performance was not an artifact of a particular data split.

D. Hyperparameter Tuning

To enhance the performance and generalization ability of the FNN model, a hyperparameter optimization process was conducted. Several key hyperparameters were selected based on their relevance to the architecture design and learning dynamics of the model. The following were explored: the number of hidden layers, number of neurons per layer, learning rate, batch size, and dropout rate. A random search strategy was employed to navigate the hyperparameter space efficiently, avoiding the computational burden of exhaustive grid search. For each candidate configuration, the model was trained using the training set (70%) and validated on a held-out validation set (15%). The MSE on the validation set was used as the objective metric for selecting the optimal configuration. This process was repeated over multiple random combinations to identify the most performant model.

The table below summarizes the final hyperparameters that yielded the best validation performance:

TABLE II. HYPERPARAMETER SEARCH SPACE AND OPTIMAL CONFIGURATION FOR THE FNN MODEL.

Hyperparameter	Search Space	Optimal Value
Number of Hidden Layers	{1, 2, 3, 4}	3
Number of Neurons per Layer	{32, 64, 128}	64
Learning Rate	{0.001, 0.0005, 0.0001}	0.0005
Batch Size	{32, 64, 128}	64
Dropout Rate	{0.0, 0.2, 0.5}	0.2
Activation Function	{ReLU, Tanh}	ReLU
Optimizer	{Adam, RMSprop}	Adam
Loss Function	{MSE, MAE}	Mean Squared Error (MSE)

This configuration was subsequently used for training the final model, which was evaluated on the test set (15%) to assess generalization performance.

E. Cross-Validation Strategy

To ensure the robustness and generalizability of the proposed FNN, a 5-fold stratified cross-validation was conducted in addition to the standard 70-15-15 holdout split. This strategy was chosen to attenuate overfitting risks, particularly given the moderate dataset size of 1,100 records. Stratification was applied based on depression severity levels, ensuring that each fold maintained a representative distribution of depression classes.

For each iteration, 80% of the data was used for training and 20% for validation. The model was trained using the same hyperparameter configuration as defined in Table II. Performance was evaluated using three key regression metrics: MAE, RMSE, and the Coefficient of Determination R^2. These metrics were averaged across the five folds, and their standard deviations were computed to assess consistency.

F. Computational Environment

The Monte Carlo simulations and the training of the FNN were conducted within the Google Colaboratory platform, which provides a cloud-based environment equipped with hardware accelerators. The experiments were implemented using Python 3.10, leveraging prominent scientific and ML libraries including TensorFlow 2.13, Keras, scikit-learn, NumPy, and Pandas. Model training was performed on a virtual machine provisioned with an NVIDIA Tesla T4 GPU (16 GB VRAM), enabling efficient handling of computational demands associated with neural network training and cross-validation.

G. Model Description

FNN was used to predict depression levels, which offers the flexibility to capture complex non-linear relationships between the variables. The architecture of the model is defined as follows:

1) Input layer: The input layer contains n features, where n is the number of preprocessed variables.
   1) Hidden layers: Three hidden layers were added, each with 64 neurons. The activation function for each hidden layer was $ReLU$ (Rectified Linear Unit), defined as:
   $$f(x) = max(0, x) \tag{5}$$
   This activation function ensures non-linearity while avoiding the vanishing gradient problem during backpropagation.
2) Output layer: The output layer predicts the DEPRESSION score, a continuous variable, with a linear activation function. The predicted depression score is represented as:
   $$\hat{y} = W_L^T h_L + b_L \tag{6}$$
   where, W_L is the weight matrix, h_L is the activation from the last hidden layer, and b_L is the bias.
3) Loss function: The loss function used is Mean Squared Error (MSE), given by:
   $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{7}$$
   where, y_i is the true depression score, and $\hat{y_i}$ is the predicted score.
4) Optimizer: The Adam optimizer was used for gradient-based optimization. This optimizer was chosen for its efficiency in handling sparse gradients and its ability to converge quickly in neural networks. The learning rate was set to $\alpha = 0.0005$.
5) Regularization and training: A dropout rate of 0.2 was applied between layers to prevent overfitting, and the batch size was set to 64 during training.

H. Model Explainability

To improve the interpretability of the proposed model, were employed SHapley Additive exPlanations (SHAP), a model-agnostic method rooted in cooperative game theory. SHAP assigns each feature an importance value representing its marginal contribution to the model's prediction. The SHAP value ϕ_i for feature i is mathematically defined as:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right] \tag{8}$$

where, F is the set of all input features, S is a subset of F not containing feature i, $f_S(x_S)$ is the model's prediction using only the features in subset S, and $|F|$ is the total number of input features. The SHAP framework ensures a fair distribution of contributions by averaging over all possible permutations of feature subsets. In this study, was applied the `KernelExplainer` from the SHAP Python library to approximate the Shapley values, enabling both global and local interpretability of the FNN's depression prediction outcomes.

I. Hyperparameter Configuration of Benchmark Models

To ensure reproducibility of the proposed framework, Table III summarizes the hyperparameter configurations used for each ML model evaluated in this study. Default values were retained for models that did not undergo hyperparameter tuning, while the configurations for ensemble and neural network models were selected based on performance optimization through empirical testing.

It is worth noting that the two most influential predictors identified by SHAP—suicide attempt history and living situation—are categorical, non-actionable risk markers rather than modifiable intervention targets. The Monte Carlo simulations instead focus on anxiety state, economic income, and education, which were deliberately selected because they are continuous or ordinal, can be meaningfully varied across a range of values, and correspond to factors that public health policy can realistically influence. The SHAP and simulation analyses are therefore complementary rather than contradictory: SHAP identifies which factors drive individual predictions most strongly, while the simulations are restricted to the subset of factors that are both quantifiable on a continuum and actionable for intervention design.

These configurations balance computational efficiency and generalization performance. The GRU model, in particular, was trained using early stopping with a patience of 10 epochs to prevent overfitting.

TABLE III. HYPERPARAMETER CONFIGURATIONS OF THE IMPLEMENTED MACHINE LEARNING MODELS.

Model	Hyperparameters
SVR	`kernel='rbf', degree=3, gamma='scale', C=1.0, epsilon=0.1, verbose=False, max_iter=-1`
Linear Regression	`fit_intercept=True, positive=False, n_jobs=-1.`
Decision Tree Regressor	`criterion='squared_error', max_depth=10, min_samples_split=4, min_samples_leaf=2, random_state=42.`
Random Forest Regressor	`n_estimators=100, criterion='squared_error', max_depth=15, min_samples_split=4, min_samples_leaf=2, max_features='sqrt', bootstrap=True, oob_score=True, n_jobs=-1, random_state=42.`
GRU	Architecture: `GRU(80, activation='relu')` → `Dense(48, activation='relu')` → `Dense(1);` Optimizer: `Adam`, Loss: `MSE`.

J. Monte Carlo Simulation

Monte Carlo simulations were employed to simulate the effect of varying key features (e.g., ANXIETY STATE, ECONOMIC INCOME, and EDUCATION) on predicted depression scores. Monte Carlo methods allow for stochastic exploration of how changes in these variables influence the outcome, providing insight into possible future states under different scenarios. Let $X = x_1, x_2, ..., x_n$ represent the features, and let $P(\theta|X)$ represent the model's predicted depression score based on a set of features. The Monte Carlo method simulates possible future states by repeatedly sampling from the distribution of features θ, generating depression predictions:

$$\hat{y}_i = f(X, \theta) + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{9}$$

where, ϵ represents noise, and θ represents the set of variables that are varied in each simulation. The simulations were run 10,000 times to generate a distribution of predicted depression scores for different levels of ANXIETY STATE, ECONOMIC INCOME, and EDUCATION. The Monte Carlo simulation enables the estimation of the mean predicted depression score and the standard deviation for each scenario.

K. Evaluation metrics

The performance of the ML model was evaluated using standard regression metrics, while the Monte Carlo simulations were assessed using statistical measures of the depression predictions.

1) Metrics for ML models:

a) Mean Squared Error (MSE): This metric measures the average squared difference between predicted and actual depression scores, providing a sense of the overall accuracy of the model. Lower values indicate better performance.

b) Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values, providing a more interpretable measure of model accuracy. It is less sensitive to outliers than MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

c) Root Mean Squared Error (RMSE): It is defined as the square root of MSE and is useful for understanding the magnitude of prediction errors in the same units as the depression scores. Lower$RMSE$ indicates better performance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

d) R-squared (R^2): This metric indicates the proportion of variance in the depression scores explained by the model. A R^2 value close to 1 indicates a well-fitting model, while values closer to 0 suggest poor performance.

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

where, n is the total number of data points or observations in the dataset. y_i is the actual observed value of the dependent variable (in this case, the depression score) for the i^{th} data point. $\hat{y}_i$ the predicted value of the dependent variable generated by the model for the i^{th} data point. And $\bar{y}$ is the mean (average) of the observed values of the dependent variable across all data points.

2) Metrics for Monte Carlo simulations: For the Monte Carlo simulations, was examined the distribution of predicted depression scores under various scenarios. The following metrics were used:

a) Mean predicted depression: The average predicted depression score across all simulations was calculated to estimate the central tendency of depression outcomes under different conditions:

$$\mu_{Depression} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \tag{13}$$

b) Standard deviation of predicted depression: It was used to assess the variability in predictions under each scenario. Higher standard deviations indicate more variability in depression outcomes due to uncertainty or sensitivity to certain features (e.g., anxiety state or economic income):

$$\sigma_{Depression} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \mu_{Depression})^2} \tag{14}$$

These metrics were essential for evaluating both the baseline performance of the ML model and the potential range of depression outcomes under different socio-economic and psychological conditions in the Monte Carlo simulations.

IV. RESULTS AND ANALYSIS

In this section, we present the findings from the trained ML model and the Monte Carlo simulations, which were designed to explore how key socio-economic and psychological variables influence depression predictions. The results focus on the post-COVID quarantine period in Argentina, using anxiety

state, economic income, and education levels as key predictors. These results offer insights into how depression rates may evolve in response to changes in these variables.

A. FNNs Results

The neural network model was trained to predict depression scores based on the dataset from the post-COVID quarantine period. The model's performance was evaluated using standard regression metrics, including MSE, MAE, $RMSE$, and R^2. The model's performance was compared against various benchmark models, including SVR, GRU, and simpler and more interpretable models such as LR, DT, and RF. The following table summarizes the MAE, MSE, $RMSE$, and R^2 for each model:

The benchmark models in Table IV were selected to represent three complementary methodological families relevant to this regression task: a classical statistical baseline (Linear Regression), tree-based learners of increasing complexity (Decision Tree and Random Forest), a kernel-based method commonly used in prior depression-prediction studies (SVR), and a recurrent deep learning architecture suited to sequential or tabular data (GRU). This selection allows the proposed FNN to be benchmarked not only against simpler, more interpretable models but also against more complex architectures, enabling a balanced assessment of the trade-off between predictive accuracy and model interpretability.

Beyond numerical performance metrics, the model outputs have concrete implications for individuals and communities. For example, a predicted depression score increase of even 2–3 points (as observed in high-anxiety simulations) could correspond to a transition from mild to moderate depression severity, potentially requiring clinical intervention. Similarly, the reduction in depression scores associated with higher income levels suggests that improved financial stability could translate into better access to healthcare, reduced chronic stress, and enhanced quality of life. These patterns are consistent with observed public health trends during the pandemic in Argentina, where economic instability and social isolation amplified mental health challenges.

Fig. 2 shows the actual depression values with the FNN model's predicted values. This visualization demonstrates the accuracy of the model in capturing depression trends.

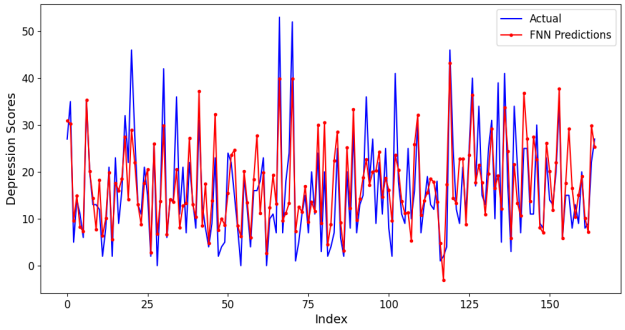


Fig. 2. Actual vs. predicted depression values.

The FNN model outperforms SVR, GRU, DT, RF in terms of MAE, MSE, $RMSE$, and R^2, highlighting its ability to capture complex, non-linear patterns in the data. Interestingly, LR achieved the highest R^2 (0.693) and lowest RMSE (6.25) among all models, suggesting that the underlying relationships between predictors and depression scores may be partially linear. The RF model also performed strongly, indicating the value of ensemble tree-based approaches in this context. In contrast, the DT model displayed the weakest performance, with the highest error metrics and lowest R^2 (0.435), reflecting its limited ability to generalize from the available data.

The results show that while simpler models like LR can match or even exceed the predictive accuracy of more complex architectures in certain cases, the FNN remains a robust choice for scenarios where capturing subtle nonlinear dependencies is critical. The model's predictions reinforce that higher anxiety levels, lower economic income, and lower education are strong predictors of elevated depression scores, aligning with previous findings in the literature [27].

Table V presents the mean and standard deviation of the evaluation metrics (MSE, MAE, $RMSE$, and R^2) obtained through the 5-fold stratified cross-validation procedure for all benchmark models. This evaluation provides a more comprehensive measure of model robustness and generalization compared to a single holdout validation, as it accounts for variability across different data partitions.

The results from the 5-fold stratified cross-validation confirm the robustness and consistency of the evaluated models. The FNN maintained superior performance, achieving the lowest average errors (MSE = 49.84 ± 6.07, MAE = 5.53 ± 0.30, RMSE = 7.06 ± 0.43) and the highest coefficient of determination (R^2 = 0.593 ± 0.048). SVR followed closely, demonstrating strong predictive capability and stability across folds, while LR, DT, and RF showed moderate performance with higher variability, reflecting their sensitivity to data partitioning. GRU, by contrast, exhibited the weakest generalization, reaffirming its limited suitability for non-temporal data.

When comparing these results with those obtained from the single holdout evaluation in Table IV, the overall ranking of models remains consistent, The FNN and SVR still lead in accuracy and generalization. However, the cross-validation results present slightly higher error values and reduced R^2 scores, which is expected due to the more rigorous partitioning of the dataset. This consistency between the two evaluation protocols supports the reliability of the FNN model and the stability of the comparative analysis across different validation strategies.

B. Model Explainability Using SHAP

To address the interpretability limitations of the FNN, were employed SHapley Additive exPlanations (SHAP)—a game-theoretic approach that quantifies the contribution of each input variable to the model's predictions. This method provides both local and global explanations by assigning SHAP values to each feature for individual predictions and aggregating them to reveal global importance.

Fig. 3 illustrates the SHAP summary plot, where each dot represents a SHAP value for a feature across different individuals. The horizontal axis reflects the magnitude and direction of each feature's impact on the predicted depression

TABLE IV. COMPARISON OF PERFORMANCE RESULTS

Model	MSE	MAE	$RMSE$	R^2
SVR	52.27	5.58	7.22	0.58
GRU	43.38	4.86	6.58	0.56
LR	39.06	4.79	6.25	0.693
DT	71.90	6.55	8.48	0.435
RF	42.52	4.91	6.52	0.666
FNN	43.38	4.72	6.32	0.64

TABLE V. PERFORMANCE COMPARISON USING 5-FOLD CROSS-VALIDATION

Model	**MSE**	**MAE**	**RMSE**	**R^2**
SVR	50.69 ± 5.00	5.74 ± 0.25	7.11 ± 0.34	0.584 ± 0.043
GRU	112.55 ± 17.63	8.36 ± 0.78	10.57 ± 0.87	0.082 ± 0.107
LR	64.54 ± 2.76	6.15 ± 0.19	8.03 ± 0.17	0.470 ± 0.032
DT	73.55 ± 8.17	6.78 ± 0.33	8.56 ± 0.48	0.397 ± 0.061
RF	73.43 ± 4.15	6.57 ± 0.20	8.57 ± 0.24	0.398 ± 0.023
FNN	$\mathbf{49.84 \pm 6.07}$	$\mathbf{5.53 \pm 0.30}$	$\mathbf{7.06 \pm 0.43}$	$\mathbf{0.593 \pm 0.048}$

scores, while the color gradient (blue to red) indicates the original feature value from low to high, respectively.

The analysis revealed that the most influential features were:

- History of suicide attempt (Yes/No): Individuals with a positive history (in red) were associated with significantly higher predicted depression scores.
- Living with somebody: Shared living conditions appeared to moderate depressive symptoms, likely reflecting the protective role of social connectedness.
- Age: Younger individuals tended to have higher SHAP values, suggesting greater vulnerability to depressive symptoms in this subgroup.

Other relevant features included suicide risk, anxiety state, and education level, each showing moderate predictive contributions. Economic income and mental disorder history had smaller but still meaningful effects.

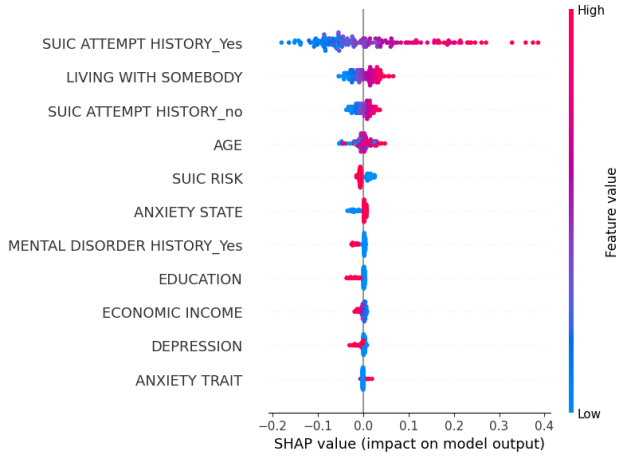


Fig. 3. SHAP summary plot showing feature importance and directional impact on depression prediction.

C. Monte Carlo Simulation Results

To further understand how variations in key factors influence depression, Monte Carlo simulations were employed. These simulations allowed for the exploration of "what-if" scenarios, simulating changes in anxiety state, economic income, and education levels.

1) Impact of anxiety state on depression: The first set of simulations explored the impact of varying anxiety state on predicted depression scores. Anxiety levels were varied from 20 to 50, reflecting low to high anxiety states. The results show a clear relationship between increasing anxiety state and rising depression scores.

$$Mean\,Predicted\,Depression = f(Anxiety\,State)$$

As anxiety state increased from 20 to 50, the mean predicted depression scores rose from 23.80 to 25.50. The standard deviation of predicted depression also increased, indicating greater variability in depression outcomes at higher anxiety states. The results suggest that anxiety state is a critical factor in determining depression levels, with higher anxiety states strongly associated with elevated depression. This finding underscores the need for interventions that target anxiety reduction as part of broader mental health support efforts. Fig. 4 shows the relationship between anxiety state and predicted depression scores. The plot clearly illustrates the upward trend in depression as anxiety state increases, with shaded error bands representing the standard deviation across the simulations.

2) Impact of economic income on depression: The second set of simulations examined how variations in economic income affect depression. Economic income was divided into three categories: low income, middle income, and high income. The Monte Carlo simulations generated predictions for depression based on these categories.

$$Mean\,Predicted\,Depression = f(Economic\,Income)$$

Low-income respondents exhibited the highest mean predicted depression score of 26.01, compared to 23.84 for middle-income and 22.73 for high-income respondents. This finding highlights the significant impact of economic hardship on mental health, with lower economic income strongly linked

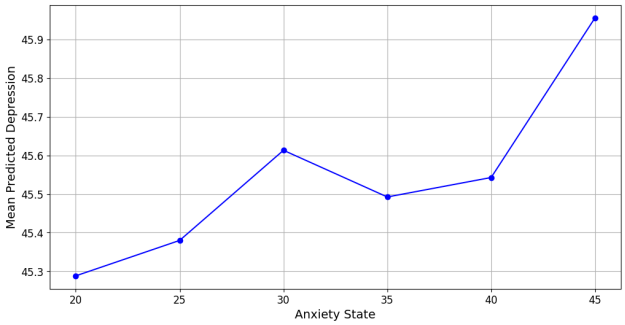


Fig. 4. Anxiety state vs. predicted depression.

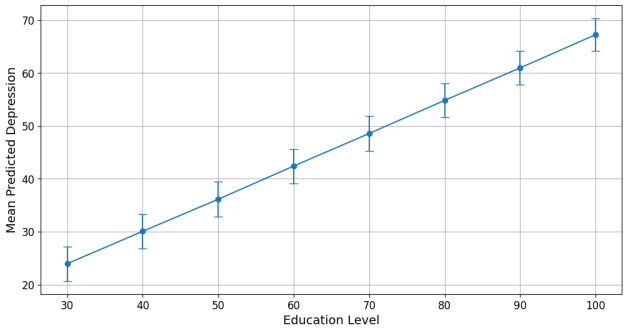


Fig. 6. Education vs. predicted depression.

to higher depression scores. Fig. 5 shows the distribution of predicted depression scores across different economic income levels. The plot clearly demonstrates that lower-income individuals are at a much higher risk of depression.

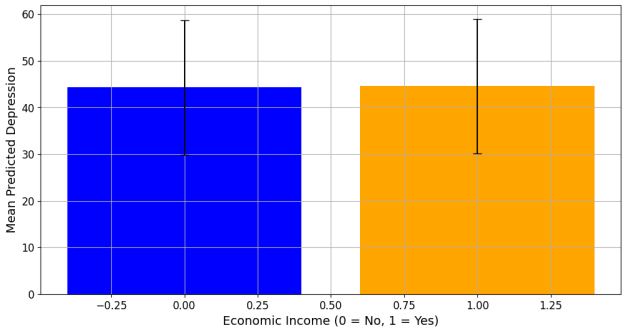


Fig. 5. Economic income vs. predicted depression.

3) Impact of education on depression: The final set of simulations explored the relationship between education levels and depression. Education was categorized into low education (incomplete elementary school), middle education (completed high school), and high education (completed university or postgraduate). The simulations provide insight into how educational attainment influences depression.

$$Mean\,Predicted\,Depression = f(Education\,Level)$$

Fig. 6 illustrates the impact of education on depression, with lower education levels associated with significantly higher depression scores. This highlights the importance of educational attainment in mitigating the effects of other mental health stressors. Respondents with low education had the highest predicted depression scores, with a mean of 25.35, compared to 23.45 for middle education and 21.98 for high education. Higher education appears to serve as a protective factor against depression, reducing vulnerability to the socio-economic stress that contribute to poor mental health.

V. Discussion

A. Interpretation of Results

The results of this study offer valuable insights into the relationship between depression and key socio-economic and psychological factors during the post-COVID-19 quarantine period in Argentina.

The anxiety state emerged as a significant predictor of depression, with higher anxiety levels consistently correlating with elevated depression scores. This finding is consistent with established mental health literature, which emphasizes that anxiety and depression often co-occur and exacerbate each other. In particular, individuals experiencing higher levels of anxiety may struggle with emotional regulation and face greater difficulty managing stress, leading to increased vulnerability to depression.

Economic income also plays a crucial role in depression outcomes. The Monte Carlo simulations demonstrated that individuals with lower economic income are more likely to experience higher depression scores. This result aligns with existing research showing that economic hardship can lead to mental health challenges, particularly in the aftermath of the pandemic. Economic instability and financial stress are known contributors to depression, as they increase uncertainty and limit access to mental health resources.

Education levels revealed a protective effect against depression. Individuals with higher educational attainment generally displayed lower depression scores, suggesting that education may serve as a buffer against socio-economic stressors. Education equips individuals with critical thinking skills, problem-solving abilities, and greater access to information, all of which can enhance resilience in the face of mental health challenges.

The cross-validation results provide crucial insights into the reliability of the FNN model. The consistency of MAE, RMSE, and R^2 scores across all five folds indicates that the model is not overly sensitive to specific data partitions. This robustness is particularly important in mental health prediction tasks, where overfitting could lead to misleading or biased recommendations. The observed stability suggests that the training configuration, feature set, and regularization strategy are effective in balancing bias and variance. The relatively high R^2 value confirms that the selected predictors—particularly anxiety state, education, income, and living situation—are strong indicators of depression severity within this dataset.

Comparing the single holdout and cross-validation results also provides additional interpretative depth. Although both evaluation protocols revealed a consistent ranking of the models - with FNN and SVR performing best, cross-validation produced slightly higher error measures, as expected due to its stricter data partitioning. This difference underscores the reliability of the reported results, confirming that the perfor-

mance of the FNN is not dependent on a specific data split and that its predictions remain stable under varying conditions.

The flexibility of supervised neural networks in capturing complex nonlinear dynamics has been demonstrated in diverse domains, further reinforcing their suitability for the mental health prediction task.

From a model performance perspective, the comparison with simpler and more interpretable models provides valuable context. Although the FNN achieved strong predictive performance, LR unexpectedly attained the highest R^2 (0.693) and lowest RMSE (6.25), suggesting that a significant portion of the relationship between predictors and depression scores can be captured by a predominantly linear model. This finding implies that for certain applications where transparency and interpretability are prioritized—such as clinical decision-making or policy formulation—LR may be a viable alternative without substantial loss in predictive accuracy. The RF model also performed competitively, underscoring the potential of ensemble-based methods to capture non-linear interactions. In contrast, the DT model showed the weakest results, highlighting the limitations of single-tree models in capturing complex and high-dimensional relationships in mental health data.

The SHAP-based interpretability results provide a clearer understanding of the FNN's predictive behavior and offer actionable implications for mental health research and policy. The prominent influence of suicide attempt history and social living conditions supports the established literature on the psychosocial determinants of depression. Moreover, the moderate effects of anxiety, income, and education suggest that mental health interventions should adopt a multidimensional approach, targeting both emotional well-being and socioeconomic stability. Importantly, this explainability layer reinforces the validity of the FNN model by aligning its internal logic with known clinical risk factors. Such transparency enhances the model's credibility and applicability in real-world mental health decision-making, especially in post-pandemic contexts where resource allocation must be both evidence-based and ethically grounded.

B. Comparison with Existing Studies

The findings from this study are consistent with the broader literature on mental health during and after the COVID-19 pandemic. Several studies have reported a strong association between anxiety and depression, particularly in populations affected by extended periods of lockdown and economic uncertainty. For instance, a study by Mogesie Necho et al. [43] found that anxiety was a major contributor to depression symptoms during the pandemic across various populations.

Similarly, the association between economic income and mental health is well-documented. Jenna M. Wilson et al. [44] reported that financial insecurity during the pandemic was a significant risk factor for developing depression. The results of this study confirm the robustness of these findings, showing that individuals with lower income are at greater risk of depression.

Regarding education, the protective role of higher education aligns with the work of other researchers, such as, Aleksander Aristovnik et al. [45] who demonstrated that individuals with higher educational levels were better able to deal with the psychological impact of the COVID-19 pandemic. The ability to access information, develop coping strategies, and seek professional help likely explains this buffering effect. While this study reaffirms much of the existing literature, the integration of ML models and Monte Carlo simulations adds a novel predictive component that enhances the understanding of how these factors interact to affect depression outcomes in the post-quarantine period.

In terms of predictive performance, the results obtained here are also comparable to those reported in related ML-based depression-prediction studies. For example, prior work applying ML models to depression or mental-health outcome prediction has typically reported R^2 values in the range of 0.4–0.7 depending on the dataset and feature set used [31], [30], [32]. The R^2 obtained by the proposed FNN (0.64 on the holdout set, 0.593 ± 0.048 under cross-validation) falls within this range, and the competitive performance of the simpler LR model ($R^2 = 0.693$) is consistent with the observation, also noted in some of these studies, that socio-economic and psychological predictors of depression often exhibit substantial linear structure. This comparison suggests that, despite differences in population (Argentina, post-quarantine) and feature set, the predictive ceiling achieved in this study aligns with the broader literature, while the addition of Monte Carlo simulation and SHAP-based explainability extends beyond what is typically reported in these comparable works.

The results of this study hold direct practical relevance for mental health policy, clinical practice, and community-based interventions. For public health policymakers, the identified strong influence of anxiety levels on depression suggests prioritizing large-scale anxiety screening and prevention programs, particularly during and after crises such as pandemics. Economic income effects indicate that targeted financial support and employment assistance programs could reduce mental health disparities in low-income populations. In the education sector, the protective role of higher educational attainment underscores the value of long-term investment in education and mental health literacy programs, which can enhance coping strategies and resilience. For clinical practitioners, integrating socio-economic indicators into mental health assessment tools can improve the early identification of high-risk individuals. These findings can also guide the allocation of limited resources by focusing on communities with combined risk factors of high anxiety, low income, and low education, thus maximizing the effectiveness of interventions in real-world settings.

From a physical and societal perspective, these statistical relationships mirror observable patterns in the Argentine post-quarantine context. Elevated anxiety levels often manifested in increased demand for mental health services, more frequent psychosomatic complaints (e.g., headaches, fatigue, sleep disturbances), and reduced workplace productivity. Economic hardship was physically reflected in food insecurity, housing instability, and restricted access to healthcare, all of which exacerbate psychological distress. In contrast, people with higher education levels often exhibited better stress management strategies, healthier lifestyle choices, and stronger social networks, tangible factors that help mitigate the risk of depression. Such direct physical manifestations reinforce

the importance of addressing these socioeconomic variables in both policy and clinical interventions.

C. Limitations

Despite the strength of the findings, several limitations must be acknowledged. First, the dataset used in this study, while comprehensive, is limited to the post-quarantine period in Argentina. This temporal and geographical limitation may affect the generalizability of the results to other regions or periods. Future studies should consider expanding the dataset to include a more diverse population and longer time spans to better capture the evolving nature of depression in the aftermath of the pandemic. Second, while the ML model demonstrated strong predictive power, there are inherent limitations in using FNN models for psychological phenomena. Depression is a multifaceted condition influenced by numerous unobserved factors, including genetic predispositions, social support, and pre-existing mental health conditions, which may not be fully captured by the available dataset. Further research could explore integrating additional psychological variables into the model for a more holistic understanding of depression. Additionally, the Monte Carlo simulations, while useful in exploring different "what-if" scenarios, rely on assumptions about the distributions of variables such as anxiety and economic income. These assumptions may not fully reflect the real-world complexities of how these variables fluctuate over time and across different populations. Further studies could refine these simulations with more precise data.

VI. Conclusion and Future Works

This study has provided a comprehensive analysis of depression in Argentina during the post-COVID-19 quarantine period, using ML models and Monte Carlo simulations to predict depression outcomes based on socio-economic and psychological factors. The key findings highlight the significant role of anxiety state, economic income, and education in shaping depression levels. The results show that individuals with higher anxiety, lower income, and lower educational attainment are more vulnerable to depression, underscoring the complexity of mental health challenges in the wake of the pandemic.

The proposed FNN achieved strong predictive performance, with $MAE = 4.72 \pm 0.30$, $MSE = 43.38 \pm 3.95$, $RMSE = 6.32 \pm 0.43$, and $R^2 = 0.64 \pm 0.048$, outperforming the SVR, GRU, DT, and RF benchmarks. LR, however, achieved the highest R^2 (0.693) and a slightly lower RMSE (6.25), suggesting that part of the relationship between predictors and depression scores may be linear, whereas DT had the weakest performance (R^2 = 0.435).

The predictive model developed in this study offers a powerful tool for mental health practitioners and policymakers. By accurately forecasting depression outcomes based on key variables, the model can help target interventions to populations most at risk. For instance, the findings on anxiety state suggest that mental health support programs should prioritize anxiety reduction strategies. Similarly, the association between economic hardship and depression underscores the need for economic policies that provide financial relief to the most vulnerable groups. Education is also shown to play a protective role, suggesting that long-term investments in education could have significant mental health benefits.

This work also highlights the trade-off between interpretability and accuracy: while deep learning models capture subtle non-linear dependencies, simpler models may be more transparent and easier to implement in certain clinical or policy contexts.

Looking ahead, there are several areas for future research. First, incorporating additional variables, such as social support, access to mental health services, and pre-existing mental health conditions, could improve the accuracy of the model and provide a more nuanced understanding of depression. Additionally, exploring the long-term mental health effects of the pandemic through longitudinal studies could offer insights into how depression evolves over time. Finally, expanding the dataset to include diverse populations and geographical regions would help generalize the findings and provide a more global perspective on post-pandemic mental health.

References

[1] J. Xiong, O. Lipsitz, F. Nasri, L. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed, and R. McIntyre, "Impact of covid-19 pandemic on mental health in the general population: A systematic review," *Journal of Affective Disorders*, vol. 277, pp. 55 – 64, 2020.

[2] C. Benke, L. K. Autenrieth, E. Asselmann, and C. Pané-Farré, "Lockdown, quarantine measures, and social distancing: Associations with depression, anxiety and distress at the beginning of the covid-19 pandemic among adults from germany," *Psychiatry Research*, vol. 293, pp. 113 462 – 113 462, 2020.

[3] A. Serafim, R. S. S. Durães, C. C. Rocca, P. D. Gonçalves, F. Saffi, A. Cappellozza, M. Paulino, R. Dumas-Diniz, S. Brissos, R. Brites, L. Alho, and F. Lotufo-Neto, "Exploratory study on the psy chological impact of covid-19 on the general brazilian population." *PloS one*, vol. 16 2, p. e0245868, 2021.

[4] J. Mari and M. Oquendo, "Mental health consequences of covid-19: the next global pandemic," *Trends in Psychiatry and Psychotherapy*, vol. 42, pp. 219 – 221, 2020.

[5] M. M. Kshirsagar, A. S. Dodamani, G. A. Dodamani, V. R. Khobragade, and R. N. Deokar, "Impact of covid-19 on mental health: An overview." *Reviews on recent clinical trials*, 2021.

[6] J. Ballivian, M. Alcaide, D. Cecchini, D. Jones, J. Abbamonte, and I. Cassetti, "Impact of covid-19-related stress and lockdown on mental health among people living with hiv in argentina." *Journal of acquired immune deficiency syndromes*, vol. 85 4, pp. 475–482, 2020.

[7] A. Rubinstein, F. Filippini, A. Santoro, A. L. L. Osornio, A. L. Bardach, E. Navarro, C. A. Cejas, S. Bauhoff, F. Augustovski, A. Pichón-Rivière, and E. L. L. Yeyati, "Lives versus livelihoods: The epidemiological, social, and economic impact of covid-19 in latin america and the caribbean." *Health affairs*, vol. 42 12, pp. 1647–1656, 2023.

[8] M. del Valle, L.-M. Hernán, G.-T. Rosario, P. F. Martín, G. M. Jonás, Y.-A. Matías, A. M. Laura, C.-J. Lorena, and U. Sebastián, "More than a year of pandemic: Longitudinal assessment of anxiety and depressive symptoms in the argentine general population during the covid-19 outbreak," *Stress and Health*, 2022.

[9] M. J. Etchevers, C. J. Garay, N. I. Putrino, N. Helmich, and G. Lunansky, "Argentinian mental health during the covid-19 pandemic: A screening study of the general population during two periods of quarantine," *Clinical Psychology in Europe*, vol. 3, 2021.

[10] M. Hossain, S. Tasnim, A. Sultana, F. Faizah, H. Mazumder, L. Zou, E. McKyer, H. U. Ahmed, and P. Ma, "Epidemiology of mental health problems in covid-19: a review," *F1000Research*, vol. 9, 2020.

[11] Y. Yin, X. Yang, L. Gao, S. Zhang, M. Qi, L. Zhang, Y. Tan, and J. Chen, "The association between social support, covid-19 exposure, and medical students' mental health," *Frontiers in Psychiatry*, vol. 12, 2021.

[12] F. Torrente, A. Yoris, D. H. Low, P. López, P. Bekinschtein, M. Cetkovich, and F. Manes, "Sooner than you think: A very early affective reaction to the covid-19 pandemic and quarantine in argentina," *Journal of Affective Disorders*, vol. 282, pp. 495 – 503, 2020.

[13] R. Qasrawi, S. V. Polo, D. A. Al-Halawa, S. Hallaq, and Z. Abdeen, "Assessment and prediction of depression and anxiety risk factors in schoolchildren: Machine learning techniques performance analysis," *JMIR Formative Research*, vol. 6, 2022.

[14] C. Bone, M. Simmonds-Buckley, R. Thwaites, D. M. Sandford, M. Merzhvynska, J. A. Rubel, A.-K. Deisenhofer, W. Lutz, and J. Delgadillo, "Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data." *The Lancet. Digital health*, vol. 3 4, pp. e231–e240, 2021.

[15] P. Kumar, S. Garg, and A. Garg, "Assessment of anxiety, depression and stress using machine learning models," *Procedia Computer Science*, vol. 171, pp. 1989–1998, 2020.

[16] H. D. A. Rahimapandi, R. Maskat, R. Musa, and N. Ardi, "Depression prediction using machine learning: a review," *IAES International Journal of Artificial Intelligence (IJ-AI)*, 2022.

[17] A. Grzenda, W. Speier, P. Siddarth, A. Pant, B. Krause-Sorio, K. Narr, and H. Lavretsky, "Machine learning prediction of treatment outcome in late-life depression," *Frontiers in Psychiatry*, vol. 12, 2021.

[18] B. Pham, T. Nguyen-Thoi, H. Ly, M. D. Nguyen, N. Al-Ansari, V. Tran, and T.-T. Le, "Extreme learning machine based prediction of soil shear strength: A sensitivity analysis using monte carlo simulations and feature backward elimination," *Sustainability*, 2020.

[19] J. Yoo and M. Rho, "Large-scale survey data analysis with penalized regression: A monte carlo simulation on missing categorical predictors," *Multivariate Behavioral Research*, vol. 57, pp. 642 – 657, 2021.

[20] C. Wang, R. Pan, X. Wan, Y. Tan, L. Xu, R. McIntyre, F. Choo, B. Tran, R. Ho, V. Sharma, and C. Ho, "A longitudinal study on the mental health of general population during the covid-19 epidemic in china," *Brain, Behavior, and Immunity*, vol. 87, pp. 40 – 48, 2020.

[21] A. R. Borracci and D. N. Giglio, "Forecasting the effect of social distancing on covid-19 autumn-winter outbreak in the metropolitan area of buenos aires." *Medicina*, vol. 80 Suppl 3, pp. 7–15, 2020.

[22] G. Serafini, B. Parmigiani, A. Amerio, A. Aguglia, L. Sher, and M. Amore, "The psychological impact of covid-19 on the mental health in the general population," *QJM: An International Journal of Medicine*, vol. 113, pp. 531 – 537, 2020.

[23] H. Badellino, M. E. Gobbo, E. Torres, and M. E. Aschieri, "Early indicators and risk factors associated with mental health problems during covid-19 quarantine: Is there a relationship with the number of confirmed cases and deaths?" *The International Journal of Social Psychiatry*, vol. 67, pp. 567 – 575, 2020.

[24] S. Dubey, P. Biswas, R. Ghosh, S. Chatterjee, M. Dubey, S. Chatterjee, D. Lahiri, and C. Lavie, "Psychosocial impact of covid-19," *Diabetes & Metabolic Syndrome*, vol. 14, pp. 779 – 788, 2020.

[25] L. Lei, X. Huang, S. Zhang, J. Yang, L. Yang, and M. Xu, "Comparison of prevalence and associated factors of anxiety and depression among people affected by versus people unaffected by quarantine during the covid-19 epidemic in southwestern china," *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, vol. 26, pp. e924 609–1 – e924 609–12, 2020.

[26] D. Witteveen and E. Velthorst, "Economic hardship and mental health complaints during covid-19," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 27 277 – 27 284, 2020.

[27] F. Torrente, A. Yoris, D. Low, P. López, P. Bekinschtein, G. Vázquez, F. Manes, and M. Cetkovich, "Psychological symptoms, mental fatigue and behavioural adherence after 72 continuous days of strict lockdown during the covid-19 pandemic in argentina," *BJPsych Open*, vol. 8, 2021.

[28] Z. Zhou, D. Luo, B. Yang, and Z. Liu, "Machine learning-based prediction models for depression symptoms among chinese healthcare workers during the early covid-19 outbreak in 2020: A cross-sectional study," *Frontiers in Psychiatry*, vol. 13, 2022.

[29] R. Aggarwal, N. Girdhar, and Alpana, "Machine learning role in cognitive mental health analysis amid covid-19 crisis: A critical study," *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1, pp. 90–96, 2022.

[30] N. Z. Bakkeli, "Predicting psychological distress during the covid-19 pandemic: Do socioeconomic factors matter?" *Social Science Computer Review*, vol. 41, pp. 1227 – 1251, 2022.

[31] R. Razavi, A. Gharipour, and M. Gharipour, "Depression screening using mobile phone usage metadata: a machine learning approach," *Journal of the American Medical Informatics Association : JAMIA*, 2020.

[32] J. Sun, R. Liao, M. Shalaginov, and T. Zeng, "A machine-learning approach for predicting depression through demographic and socioeconomic features," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2203–2207, 2022.

[33] F. N. Al-Wesabi, H. Alsolai, A. M. Hilal, M. A. Hamza, M. A. Duhayyim, and N. Negm, "Machine learning based depression, anxiety, and stress predictive model during covid-19 crisis," *Computers, Materials & Continua*, 2022.

[34] K. Hueniken, N. Somé, M. Abdelhack, G. W. Taylor, T. E. Marshall, C. M. Wickens, H. Hamilton, S. Wells, and D. Felsky, "Machine learning–based predictive modeling of anxiety and depressive symptoms during 8 months of the covid-19 global pandemic: Repeated cross-sectional survey study," *JMIR Mental Health*, vol. 8, 2021.

[35] L. Flesia, M. Monaro, C. Mazza, V. Fietta, E. Colicino, B. Segatto, and P. Roma, "Predicting perceived stress related to the covid-19 outbreak through stable psychological traits and machine learning models," *Journal of Clinical Medicine*, vol. 9, 2020.

[36] S. Maltezos and A. Georgakopoulou, "Novel approach for monte carlo simulation of the new covid-19 spread dynamics," *Infection, Genetics and Evolution*, vol. 92, pp. 104 896 – 104 896, 2020.

[37] L. E. de Sousa, P. H. O. Neto, and D. A. S. Filho, "Kinetic monte carlo model for the covid-19 epidemic: Impact of mobility restriction on a covid-19 outbreak." *Physical review. E*, vol. 102 3-1, p. 032133, 2020.

[38] G. Xie, "A novel monte carlo simulation procedure for modelling covid-19 spread over time," *Scientific Reports*, 2020.

[39] I. Ciufolini and A. Paolozzi, "An improved mathematical prediction of the time evolution of the covid-19 pandemic in italy, with a monte carlo simulation and error analyses," *European Physical Journal plus*, 2020.

[40] N. Alkhateeb, F. Sallabi, S. Harous, and M. Awad, "A study on predicting the outbreak of covid-19 in the united arab emirates: A monte carlo simulation approach," *Mathematics*, 2022.

[41] L. L. Steinmetz, "Data from: Mental health of people in argentina post-quarantine covid-19," 2021. [Online]. Available: https://rdu.unc.edu.ar/handle/11086/20168

[42] A. Siham, S. Sara, and A. Abdellah, "Feature selection based on machine learning for credit scoring : An evaluation of filter and embedded methods," in *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2021, pp. 1–6.

[43] M. Necho, M. Tsehay, M. Birkie, G. Biset, and E. Tadesse, "Prevalence of anxiety, depression, and psychological distress among the general population during the covid-19 pandemic: A systematic review and meta-analysis," *International Journal of Social Psychiatry*, vol. 67, pp. 892 – 906, 2021.

[44] J. M. Wilson, J. Lee, H. N. Fitzgerald, B. Oosterhoff, B. Sevi, and N. J. Shook, "Job insecurity and financial concern during the covid-19 pandemic are associated with worse mental health." *Journal of Occupational & Environmental Medicine*, 2020.

[45] A. Aristovnik, D. Keržič, D. Ravšelj, N. Tomaževič, and L. Umek, "Impacts of the covid-19 pandemic on life of higher education students: A global perspective," *Sustainability*, 2020.

Context-Aware Sentiment Analysis of E-Commerce Reviews Using a BERT-CNN-BiLSTM Hybrid Model

Mahmud Reza Mahim, Fahad Bin Z Islam, Md Saklain Mahmud, Md. Imtiyaz Hasan, Sifat Rahman Ahona
Department of Computer Science and Engineering
American International University-Bangladesh (AIUB), Dhaka, Bangladesh

***Abstract*—User-generated product reviews are an essential source of information in e-commerce; nevertheless, the huge volume and varying quality of review texts make extracting insights difficult. The conventional approach to sentiment classification is limited in terms of recognizing contextual and aspect-oriented sentiment clues in the text. This study proposes a hybrid architecture that uses contextual sentiment clues for e-commerce reviews sentiment classification. The experiments are conducted using FABSA dataset which contains reviews annotated in terms of multiple aspects–sentiments pairs. Every review is expanded in terms of aspect-oriented sentiment classification samples. This makes it possible to learn fine-grained polarities of sentiments associated with specific aspects of product reviews. To tackle the problem of class imbalance in the data, the experiments employ the method of stratified oversampling in combination with the use of class-weighted cross-entropy loss function. The empirical results show that the improved hybrid BERT–CNN–BiLSTM model achieves 92.20% validation accuracy, 92.25% weighted F1-score, and 86.36% macro F1-score. The most noticeable progress has been made in the case of neutral class, where the F1-score has increased by 17.1 percentage points showing the improvements in minority-class recognition and decrease of class imbalance. The architecture-level ablation study shows the superior performance of the BERT–CNN–BiLSTM model compared to its simplified versions based on BERT, BERT–CNN, and BERT–BiLSTM architectures. A contextual comparison with reported FABSA baselines suggests competitive performance, although this comparison should not be interpreted as a strict leaderboard result because the baselines were not reproduced under identical experimental settings. Overall, the findings demonstrate that combining contextual transformer representations with local convolutional features and bidirectional sequential modeling can improve class-balanced sentiment classification in aspect-expanded e-commerce review data.**

Keywords*—*Sentiment analysis; e-commerce reviews; BERT; CNN; BiLSTM

I. Introduction

A. Background and Motivation

The widespread adoption of e-commerce platforms has made user-generated product reviews one of the most influential sources of information in consumer purchase decisions [1], [2]. User-generated reviews help consumers reduce uncertainty regarding product quality, functionality, usability, and post-purchase satisfaction [2], [3]. Therefore, review information is increasingly used in recommendation and decision-support systems. However, the abundance of review content creates information overload, making it difficult for consumers and online platforms to filter useful information from a large pool of reviews. In addition, many reviews are short, generic, sentiment-heavy, or low in informational value, while detailed experience-based reviews may remain unnoticed [4], [5], [6], [7].

Traditional sentiment analysis methods generally focus on assigning broad polarity labels, such as positive, negative, or neutral, to a given text. Although such approaches are useful, they often fail to capture fine-grained contextual information related to specific product aspects, usage conditions, and user intentions [8], [9]. For example, a review may express positive sentiment toward one aspect of a product while expressing negative sentiment toward another. Similarly, short expressions, such as "Great product!", may indicate positive polarity but provide limited contextual detail about the product experience. These cases show that sentiment classification in e-commerce reviews requires models that can capture not only overall polarity, but also the contextual cues surrounding product-related aspects [6], [7], [10], [11].

Lexicon-based and conventional machine learning methods have provided useful and interpretable baselines for sentiment analysis. Some recent studies also show that calibrated lexicon-based approaches can remain competitive in specific domains [9]. However, conventional models often struggle with domain-specific language, compositional meaning, sarcasm, implicit sentiment, and the semantic relationship between product aspects and opinion expressions [8], [10]. These limitations have encouraged the use of deep learning and transformer-based methods, which are more capable of learning contextual representations directly from text.

Recent studies have increasingly adopted deep learning architectures to capture both local and global dependencies in review text. BERT provides contextualized word representations through bidirectional transformer encoding, making it effective for modeling semantic relationships in sentences [12]. Convolutional neural networks (CNNs) are effective at extracting local n-gram and phrase-level sentiment features, while bidirectional long short-term memory networks (BiLSTMs) can capture sequential dependencies from both forward and backward directions [8], [13], [3]. Hybrid architectures that combine these components can integrate contextual representation learning, local feature extraction, and sequential dependency modeling within a single framework [13], [14], [15], [16].

In this context, a hybrid BERT–CNN–BiLSTM architec-

ture offers a promising approach for context-aware sentiment classification in e-commerce reviews. BERT can generate deep contextual embeddings, CNN layers can identify local sentiment-bearing patterns, and BiLSTM layers can model bidirectional dependencies across review sequences. This combination is particularly suitable for aspect-expanded review data, where each review may contain multiple aspect–sentiment pairs. Therefore, this study focuses on developing and evaluating a BERT–CNN–BiLSTM hybrid model for sentiment classification using the FABSA dataset, with emphasis on validation performance, class-balanced evaluation, ablation analysis, and cross-dataset generalization.

B. Literature Review

1) From lexicon-based to deep learning approaches: Sentiment analysis has evolved from lexicon-based techniques and classical machine learning to deep learning approaches. Early systems relied on handcrafted sentiment lexicons and surface-level representations such as bag-of-words, n-grams, and shallow syntactic cues [8], [9]. Although these methods offered a degree of interpretability and domain independence, they struggled with domain adaptation, context-dependent polarity, and subtle semantic phenomena such as irony and implicit sentiment [8], [9], [10].

Recent studies, such as MultiLexScaled, show that aggregated and rescaled lexicon-based methods can still yield competitive and interpretable sentiment scores for large review datasets. However, these methods struggle to fully understand complex compositional statistics and non-additive interactions among sentiment indicators and product-specific attributes in single reviews [9], [10].

Advances in the field of deep learning have strengthened sentiment analysis in terms of end-to-end representation learning from raw text. Under the review dataset benchmark [8], CNN-based models automatically extract local sentiment-bearing patterns as well as phrase-level features. Additionally, recurrent models like LSTM and BiLSTM have greater capacity than traditional methods to capture long-distance dependencies and sequential dynamics. Different hybrid architectures, such as CNN + LSTM and BiLSTM with attention, make sentiment analysis more robust by leveraging the synergy between complementary feature extractors [8], [17]. Early work consistently demonstrated that hybrid deep learning models outperformed single baseline models across tasks requiring understanding of text, images, and multimodal data for sentiment analysis [8], [10].

2) Transformer and hybrid architectures for sentiment analysis: The rise of pre-trained transformers, particularly BERT, was a turning point for sentiment and emotion detection, as they enable expressive contextual representations through self-attention and large scale pre-training [10], [12]. BERT based fine-tuning has achieved strong results across product review analysis, political debate analysis, and social media monitoring, often with limited task-specific data [1], [11], [18]. More recent extensions, such as RoBERTa-BiLSTM and ontology-enhanced BERT, show that combining transformer embeddings with sequential modeling or external knowledge can yield additional gains, particularly in domain-specific and aspect-level settings [10], [11], [13]. Survey evidence also suggests that transformer based and hybrid approaches now dominate leading results, with F1-scores often exceeding 0.85 on well-annotated datasets [10].

Recent research has increasingly combined BERT with CNN and BiLSTM architectures to leverage their complementary inductive biases, with CNNs capturing local n-gram features and BiLSTMs modeling sequential dependencies. Liu *et al.* proposed a BERT–BiGRU–Softmax model for e-commerce product reviews and showed that contextual embeddings combined with gated recurrent units can outperform strong recurrent baselines [1]. Shen introduced a BERT–CNN–BiLSTM–Attention model for sentiment analysis of social media texts related to aquatic product companies and reported an F1-score above 93% [14]. Khan *et al.* combined multilingual BERT with stacked CNN–BiLSTM and attention for Urdu sentiment analysis, demonstrating the usefulness of hybrid architectures in low-resource and morphologically complex settings [15]. Using this architectural pattern, they utilized a CNN–BiLSTM–BERT hybrid model and analyzed multilingual university feedback captured through Twitter, reporting above 92% accuracy. Their results demonstrate that this architectural pattern is scalable across different domains as well as languages [16]. This approach applies to sentiment and emotion detection on handwritten text, as well as text in electronic format, suggesting that these hybrid models are highly versatile [17].

Overall, the literature shows that hybrid architectures composed of pretrained transformers layered with convolutional layers and recurrent components are more effective than their single-architecture baselines in terms of accuracy and general robustness across heterogeneous sentiment-analysis benchmarks [8], [14], [15], [13], [16].

3) Aspect-level and context-aware sentiment modeling: In e-commerce review analysis, sentiment classification becomes more informative when it moves beyond review-level polarity and considers the specific product aspects discussed in the text. A single review may contain different sentiments toward different aspects of the same product, such as positive sentiment toward battery life but negative sentiment toward price or durability. Therefore, aspect-level and context-aware sentiment modeling is important for capturing fine-grained opinion patterns in product reviews [6], [7], [19].

The FABSA dataset provides a suitable benchmark for this task because it represents user reviews through aspect–sentiment annotations, allowing each review to be analyzed in relation to the product aspects it contains [19]. This type of data is especially useful for evaluating models that need to learn sentiment polarity from contextual cues rather than relying only on surface-level sentiment words. However, aspect-expanded sentiment classification also introduces several challenges, including class imbalance, contextual ambiguity, and difficulty in recognizing minority classes such as neutral sentiment [6], [20].

Previous studies have shown that transformer-based models and hybrid deep learning architectures can improve sentiment classification by learning richer representations from review text [1], [2], [3]. BERT-based models are effective for capturing contextual semantics, while CNN and BiLSTM components can further support local phrase-level feature extraction and

bidirectional sequence modeling [14], [15], [13]. Nevertheless, the combined use of BERT, CNN, and BiLSTM for aspect-expanded sentiment classification in e-commerce reviews remains an important direction for further investigation.

This gap motivates the present study, which develops and evaluates a hybrid BERT-CNN-BiLSTM model for context-aware sentiment classification on the FABSA dataset. The study focuses on validation performance, class-balanced evaluation, architecture-level ablation, and cross-dataset generalization. By doing so, it aligns the implemented experimental scope with the stated objective of improving sentiment classification in aspect-expanded e-commerce review data.

C. Research Aims and Contributions

The primary aim of this study is to design and evaluate a hybrid BERT–CNN–BiLSTM model for context-aware sentiment analysis and demonstrate that the hybrid formulation is suitable for aspect-expanded review data. More specifically, the study investigates whether combining transformer-based contextual encoding with convolutional and bidirectional sequential modeling can improve sentiment classification on aspect-expanded review instances derived from the FABSA dataset [1], [2], [14], [15], [13], [19]. In the proposed architecture, BERT provides contextual embeddings, CNN layers capture local phrase-level features, and BiLSTM layers model sequential dependencies across the review text [1], [14], [15]. This combination is intended to capture both localized sentiment cues and broader contextual structure within user reviews.

1) Research objectives:

O1: **Hybrid Model Design:** Develop a BERT–CNN–BiLSTM architecture for sentiment classification on aspect-expanded e-commerce review data, using BERT for contextual encoding, CNN for local feature extraction, and BiLSTM for bidirectional sequence modeling [1], [14], [15], [13].

O2: **Validation-Based Performance Evaluation:** Evaluate the proposed model on the FABSA dataset using standard classification metrics, including accuracy, weighted F1- score, macro F1-score, class-wise precision, recall, F1- score, and confusion-matrix analysis [19], [6], [20].

O3: **Comparative Performance Analysis:** Compare the optimized hybrid configuration with the baseline training configuration and contextualize its performance against reported results from existing sentiment-analysis models in the literature [1], [2], [3], [14].

2) Research contributions:

C1: **Hybrid BERT–CNN–BiLSTM Architecture for Sentiment Classification.** This study develops and implements a hybrid BERT–CNN–BiLSTM architecture for context-aware sentiment classification in e-commerce reviews. In the proposed model, BERT is used for contextual representation learning, CNN layers capture local sentiment-bearing phrase patterns, and BiLSTM layers model bidirectional sequential dependencies [1], [14], [15].

C2: **Aspect-Expanded Evaluation on the FABSA Dataset.** The study evaluates the proposed model on the FABSA dataset by converting review-level aspect–sentiment annotations into aspect-expanded sentiment classification instances. This allows the model to be assessed on fine-grained sentiment labels associated with product-review aspects [19].

C3: **Empirical Performance Analysis Using Class-Balanced Metrics.** The proposed model is evaluated using accuracy, weighted F1-score, macro F1-score, class-wise precision, recall, F1-score, and confusion-matrix analysis. The optimized configuration improves macro F1-score and minority-class recognition, particularly for the neutral class, indicating more balanced sentiment classification performance [6], [20].

C4: **Architecture-Level Ablation Study.** An ablation study is conducted to compare BERT, BERT–CNN, BERT–BiLSTM, and the full BERT–CNN–BiLSTM architecture under the same experimental setting. The results show that the full hybrid model achieves the strongest performance, demonstrating the complementary contribution of convolutional and recurrent components.

C5: **Cross-Dataset Generalization Analysis.** The study further examines the generalization behavior of the trained model through zero-shot evaluation on the SemEval-2014 Restaurant dataset. This analysis highlights the model's ability to transfer polarity recognition across domains while also revealing the domain-sensitive nature of neutral sentiment classification.

3) Research questions: In line with the implemented scope of the study, the following research questions guide the empirical investigation:

RQ1: How effectively can a hybrid BERT–CNN–BiLSTM model classify sentiment in aspect-expanded e-commerce reviews [1], [14], [15], [13]?

RQ2: To what extent does optimizing the hybrid training configuration improve overall and class-balanced sentiment performance compared with the initial baseline configuration [6], [20]?

RQ3: How does the proposed hybrid model compare, in contextual terms, with reported results from existing machine learning, deep learning, and transformer-based sentiment-analysis models on the FABSA benchmark [19], [1], [2], [3]?

By addressing these questions, the study seeks to improve understanding of how hybrid transformer–CNN–BiLSTM architectures can be applied to sentiment classification in e-commerce review analysis. [1], [2], [6], [14], [15], [21].

The remainder of this study is organized as follows: Section II presents the dataset, preprocessing steps, proposed BERT–CNN–BiLSTM architecture, training configuration, and evaluation metrics. Section III reports the experimental results, including validation performance, ablation analysis, class-wise evaluation, contextual comparison with reported FABSA

baselines, and cross-dataset validation. Section IV discusses the findings and interprets the strengths and limitations of the proposed model. Section V presents the limitations of the study. Finally, Section VI and Section VII provide the conclusion and future research directions, respectively.

II. Methodology

A. Dataset and Data-Preprocessing

1) Dataset: Experiments utilize the FABSA (Fine-grained Aspect-Based Sentiment Analysis) dataset, obtained from the Hugging Face Datasets Hub (`jordiclive/FABSA`). This dataset comprises product review texts accompanied by structured labels containing aspect–sentiment pairs, where each label specifies a product aspect (such as *battery*, *screen*, or *price*) and its corresponding sentiment polarity (*positive*, *negative*, or *neutral*). The dataset is pre-divided into training, validation, and test splits, which are maintained throughout all experiments to ensure fair and reproducible evaluation.

2) Data pre-processing pipeline: The raw dataset is subjected to a multi-stage pre-processing pipeline prior to input into the hybrid model.

a) Aspect–sentiment extraction: Every instance in the FABSA corpus consists of free-form reviews and labels containing pairs of aspects and sentiments. In the process of pre-processing, the reviews and their respective aspect and sentiment pairs are extracted from all the samples in the corpus. The result is a structured format with the review text in each instance and (aspect, sentiment) pairs for that review.

b) Class imbalance analysis: Before resampling, the frequency of each sentiment class (*positive*, *negative*, *neutral*) was calculated for the entire dataset, which consisted of the training, validation, and test data. The outcomes highlight a significant imbalance in class representation, with a disproportionately small number of neutral samples compared to the dominant positive class.

c) Class imbalance correction via upsampling: Class imbalance became evident in the training dataset. To address this issue, the dataset was divided into three sentiment-based groups. Stratified random sampling with replacement was then used to up-sample the minority neutral group, increasing its number to be almost equal to the number of items in the majority positive group. No changes were made to the negative group.

d) Dataset expansion: The reviews can not be mapped directly to their respective labels (which include the target words), since one review may contain more than one aspect-sentiment pairs; hence, multi-label learning is necessary. The (review, aspect, sentiment) tuple is considered as one instance, and the generation of new instances comes from the tuples to add another layer of granularity in the learning task. In particular, for each aspect-based review having k number of aspects, there will be k copies (one for each aspect) together with its respective sentiment label.

e) 80–20 Split alignment: After aspect-based expansion, the size of the augmented training set became disproportionately larger than the expanded test set. To maintain an approximate 80:20 train-to-test ratio, a uniform subsample was drawn from the augmented training set without replacement, making the final training set four times larger than the expanded test set. The validation set was kept unchanged throughout this process.

f) Tokenization: Tokenization for all review texts is performed through the BERT-base (uncased) WordPiece tokenizer. Every sequence out of the three is padded to ensure that they satisfy the condition of having a maximum sequence length as per the batch and truncated to 512 tokens, which is the total number of positional embeddings that characterize the BERT architecture. The tokenizer produces token indices and attention masks, where the attention mask distinguishes actual tokens from padding tokens.

g) Label encoding: The three sentiment polarities are mapped to integer indices: *negative* $\rightarrow$ 0, *neutral* $\rightarrow$ 1, and *positive* $\rightarrow$ 2. This encoding is applied consistently across all splits and is utilized by the cross-entropy loss function during training.

h) Class weight computation: In order to reduce the imbalance in the dataset even after the use of upsampling, class weights are calculated using the inverse frequency formula:

$$w_c = \frac{N}{N_c} \tag{1}$$

where, N denotes the total number of training instances and N_c represents the number of instances belonging to class c. Hence, in this case, the assigned weights are 1.5 (negative), 2.0 (neutral), and 1.0 (positive).

i) Dataset and data loader construction: The tokenized sequence and its corresponding label in integer format are arranged in a structured dataset, which provides individual tensor samples upon request. The training, validation, and testing sets each had their own data iterator created using a batch size of 8. For the training data iterator, randomness is applied to shuffle the data for every epoch to avoid any bias due to order, while the validation and testing data iterators follow a sequential flow.

B. Base Models

1) BERT: Bidirectional Encoder Representations from Transformers (BERT) refers to a pre-trained language model that considers both left and right-contextual information in interpreting text data [12]. The model has a multi-layered bidirectional Transformer encoder, which can be of different sizes, like $\text{BERT}_{\text{BASE}}$ (12 layers, 768 hidden units, 12 self-attention heads, 110 million parameters) and $\text{BERT}_{\text{LARGE}}$ (24 layers, 1024 hidden units, 16 self-attention heads, 340 million parameters). The inputs are fed into the network using WordPiece tokenization, where each input consists of a vocabulary of 30,000 words and includes special tokens like [CLS], used for classification tasks, and [SEP], used to differentiate sentences. Token, segment, and positional embeddings are used in the model. The pre-training phase incorporates two strategies. Masked Language Modeling (MLM) hides 15% of tokens and tasks the model with predicting them. Meanwhile, Next Sentence Prediction (NSP) helps the model to recognize sentence interactions. Fine-tuning involves task-specific training and parameter tuning. Fig. 1 presents the BERT model architecture.

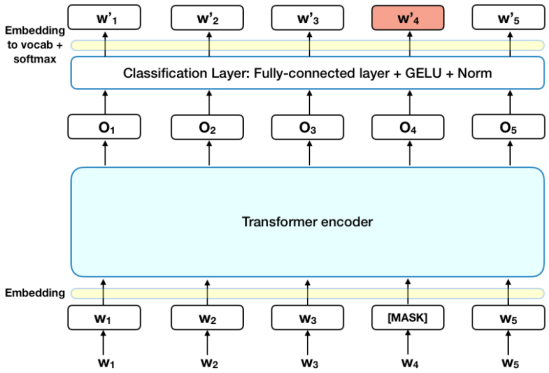


Fig. 1. BERT model architecture.

2) CNN: The process of text classification using CNNs involves the application of convolutions on the vectors of words in the sentence [22]. In this regard, each sentence is viewed as a matrix of word vectors where each word can be defined as a k-dimensional vector, hence an n x k matrix. Various filters of different sizes, for example, 3, 4, or 5, are used to produce feature maps by sliding over the matrix. In doing so, the filter uses activation functions such as ReLU or tanh. The next step involves performing the max-over-time pooling function on the obtained feature maps in order to select the most important features while at the same time allowing for the variation in sentence length. All the extracted features from various filters are concatenated to obtain a vector that is used as input to the fully connected softmax layer for classification purposes. Some models may use several channels with predetermined and trainable embeddings (see Fig. 2).

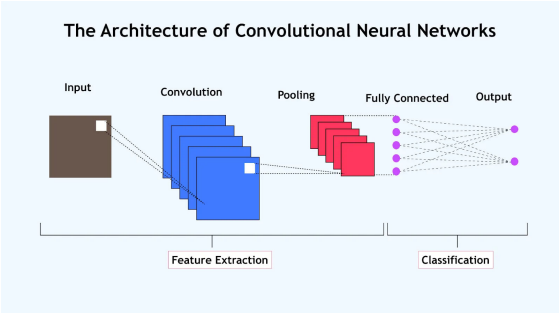


Fig. 2. CNN architecture for text classification.

3) Bi-LSTM: The Bidirectional Long Short Term Memory (Bi-LSTM) network refers to an advancement over the traditional LSTM model by processing the sequence in the forward and backward direction, thereby enabling context from past as well as future tokens [23]. It consists of two LSTM layers, where one layer processes the sequence from left to right and the other one processes it from right to left. Each LSTM unit includes three gates (input, forget, output) that regulate information retention over time. At every time step, the network concatenates the outputs from both layers. In Natural Language Processing (NLP), the Bi-LSTM models can be used for applications like sequence tagging, sentiment analysis, and text classification since they handle dependencies in both directions (see Fig. 3).

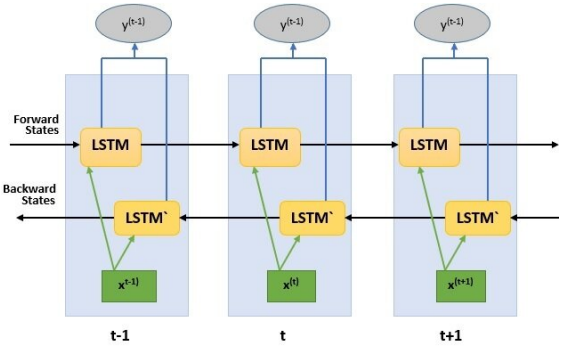


Fig. 3. Bi-LSTM architecture.

C. Hybrid Model Architecture

The proposed hybrid model consists of BERT to generate contextual embeddings, a CNN to extract local features, and a Bi-directional LSTM (Bi-LSTM) to capture sequences. These components are used for conducting context-based sentiment analysis of reviews. The generated BERT representations are passed through two modeling stages. First, three parallel CNN branches with kernel sizes of 3, 4, and 5 extract local n-gram sentiment features from the contextual token sequence. The CNN outputs are passed through ReLU activation and layer normalization before being concatenated along the feature dimension. The concatenated representation is then passed to a bidirectional LSTM to model sequential dependencies in both forward and backward directions. Finally, global max-pooling is applied over the sequence dimension, followed by dropout and a linear classification layer for three-class sentiment prediction (see Fig. 4).

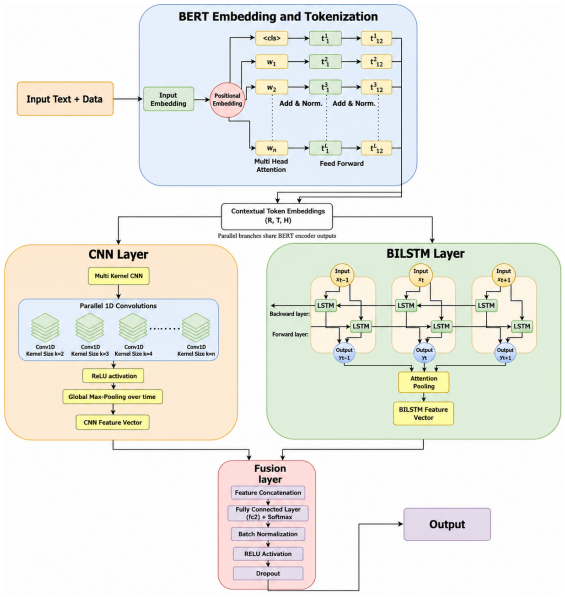


Fig. 4. Hybrid BERT-CNN-BiLSTM model architecture.

The following hyperparameters define the final implemented architecture. The CNN branch consists of three parallel convolutional layers (`Conv1d`) applied to the BERT last-hidden-state sequence (`bert-base-uncased`, hidden

dimension 768), each producing 256 output feature maps (`out_channels` $= 256$) with kernel sizes of 3, 4, and 5 tokens respectively, with padding values of 1, 2, and 2 to maintain compatible output lengths. Each convolutional output is passed through a ReLU activation followed by a `LayerNorm` layer before being fed into the shared BiLSTM. The three branch outputs—each of shape [batch $\times$ seq_len $\times$ 256]—are concatenated along the feature axis, giving a BiLSTM input size of 768 ($= 256 \times 3$). The BiLSTM has a hidden size of 128 units per direction (`bidirectional = True`), yielding a 256-dimensional output per token. Global max-pooling is then applied over the sequence dimension, followed by a dropout layer ($p = 0.5$) and a linear classification head projecting to three sentiment classes. Table I summarises these values.

TABLE I. HYPERPARAMETER SUMMARY OF THE BERT–CNN–BILSTM HYBRID MODEL.

Parameter	Value
BERT encoder	`bert-base-uncased`
BERT hidden dimension	768
Number of CNN branches	3
CNN output channels	256 per branch
CNN kernel sizes	3, 4, and 5
CNN padding	1, 2, and 2
CNN activation	ReLU
Normalization	LayerNorm after each CNN branch
Concatenated CNN dimension	768 (256×3)
BiLSTM input size	768
BiLSTM hidden units	128 per direction
BiLSTM direction	Bidirectional
BiLSTM output dimension	256
BiLSTM layers	1
Pooling method	Global max-pooling
Dropout rate	0.5
Classifier output	Linear layer ($256 \rightarrow 3$)
Batch size	8
Learning rate	1×10^{-5}
Training epochs	20
BERT fine-tuning	Frozen for first 5 epochs
Loss function	Weighted cross-entropy
Class weights	Negative: 1.5, Neutral: 2.0, Positive: 1.0

D. Model Evaluation Metrics

The performance of the proposed hybrid model is evaluated using standard classification metrics for multi-class sentiment analysis. Since the task involves three sentiment classes, namely negative, neutral, and positive, the evaluation considers both overall performance and class-wise effectiveness. The main metrics used in this study are Accuracy, Precision, Recall, F1-score, Macro-F1, Weighted-F1, and confusion-matrix analysis [20].

Accuracy measures the proportion of correctly classified samples among all samples:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2}$$

For each sentiment class, Precision measures the proportion of correctly predicted samples among all samples predicted as that class. Recall measures the proportion of correctly predicted samples among all actual samples of that class. The F1-score represents the harmonic mean of Precision and Recall. These metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. In the multi-class setting, these values are calculated for each sentiment class using a one-versus-rest interpretation.

Because sentiment datasets are often imbalanced, especially when one class has fewer samples than the others, macro-averaged and weighted-averaged metrics are also used. Macro-averaged metrics assign equal importance to each class by taking the simple average of the class-wise scores:

$$\text{Macro-Precision} = \frac{1}{C} \sum_{i=1}^{C} \text{Precision}_i \tag{6}$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{i=1}^{C} \text{Recall}_i \tag{7}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \tag{8}$$

where, C denotes the total number of sentiment classes.

Weighted-F1 is also reported to account for class distribution by weighting each class-specific F1-score according to its support:

$$\text{Weighted-F1} = \sum_{i=1}^{C} \frac{n_i}{N} \times \text{F1}_i \tag{9}$$

where, n_i is the number of samples in class i, and N is the total number of samples.

In addition to these numerical metrics, confusion matrices are used to analyze class-wise prediction behavior. The confusion matrix helps identify which sentiment classes are correctly classified and which classes are more frequently confused with one another. This is particularly important for evaluating the model's performance on minority classes such as neutral sentiment.

III. RESULTS AND ANALYSIS

In this section, the empirical results are discussed, based on the findings in the experimental setup related to the proposed BERT-CNN-BiLSTM hybrid model. In the implemented experiment, each review is expanded into aspect-text pairs, and the hybrid model is trained based on this aspect-wise partitioning approach. The results can thus be seen as part of the sentiment classification portion of this study.

A. Overall Validation Performance

Two configurations were examined with a normal hybrid model and the optimized hybrid model. As shown in Table II, optimization improved all headline validation metrics. Validation accuracy increased from 88.75% to 92.20%, weighted F1 increased from 89.32% to 92.25%, and macro F1 increased from 78.64% to 86.36% (see Fig. 5). The improvement in macro F1 is especially significant, because it indicates that the opti- mized configuration improved class balance and minority-class recognition rather than only strengthening performance on the dominant class. in the optimized configuration it dropped to 5.89 percentage points. This narrowing suggests that the optimized model produced more balanced decisions across the three sentiment classes.

TABLE II. VALIDATION PERFORMANCE OF THE ORIGINAL AND OPTIMIZED HYBRID MODELS

Configuration	Accuracy	Weighted F1	Macro F1
Original hybrid model	0.8875	0.8932	0.7864
Optimized hybrid model	0.9220	0.9225	0.8636

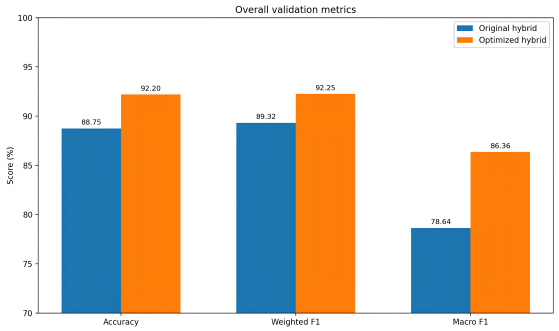


Fig. 5. Bar-chart comparison of overall validation metrics for the original and optimized hybrid models.

One of the explanations for such an increase is to examine the difference between the weighted F1 and macro F1 scores. Specifically, in the initial setting, the value of this difference was 10.68%, while after optimization, it dropped down to 5.89%. It shows that in the optimized setting, decisions on each class were more evenly distributed.

B. Architecture-Level Ablation Experiments

In order to make sure that the final performance gain achieved is due to the full hybrid architecture and not just BERT, an ablation study has been performed. Four models have been compared with the help of the exact same preprocessing approach and the exact same held out FABSA testing dataset, namely: 1) BERT, 2) BERT+CNN, 3) BERT+BiLSTM, and 4) BERT-CNN-BiLSTM. For all these tests, F1-score means weighted average test F1 from the final comparison.

According to the ablation experiments' findings, the best F1 value for testing data belongs to the complete BERT-CNN-BiLSTM architecture, whose F1 is 0.9286. On the other hand, BERT+CNN and BERT+BiLSTM models give 0.9196 and 0.9182 F1 values, respectively, while using only BERT gives an F1 value of 0.9167. Therefore, the hybrid model improved over the strongest reduced variant, BERT+CNN, by approximately 0.0090 absolute F1. The improvement is modest but consistent, suggesting that the convolutional and recurrent branches contribute complementary information as CNN layers capture local sentiment-bearing n-gram features, while the BiLSTM branch models bidirectional sequential dependencies across the review text (see Table III).

TABLE III. ABLATION EXPERIMENT RESULTS ON THE FABSA TEST SET

Model	F1-score
BERT	0.9167
BERT+CNN	0.9196
BERT+BiLSTM	0.9182
BERT-CNN-BiLSTM model	0.9286

These findings confirm the proposed architecture. While BERT provides strong context embeddings, the best F1 value is achieved only if BERT vectors are used together with the CNN and BiLSTM modules. Hence, the results support the effectiveness of the proposed architecture.

C. Learning Behavior and Convergence

In comparison, the improved architecture is able to outperform the initial configuration significantly. While the base hybrid neural network with hyperparameter tuning continues to improve in performance throughout the 20 epochs, albeit more slowly, to yield a weighted F1 score of 79.80%, the optimized neural network starts off on an advantageous note and scores over 90% validation weighted F1 from epoch 3 onwards (as shown in Fig. 6), with its peak being reached at epoch 10 at 92.68% accuracy, 92.76% weighted F1, and 0.2423 validation loss. From epoch 10 onward, the validation loss slightly increases compared to previous epochs, whereas training accuracy keeps improving, signaling moderate overfitting towards the end of the epochs. Considering the separation variable analysis approach, it can be concluded that early stopping at epoch 10 would have been optimal to retain maximal generalization efficiency while cutting down on training expenses.

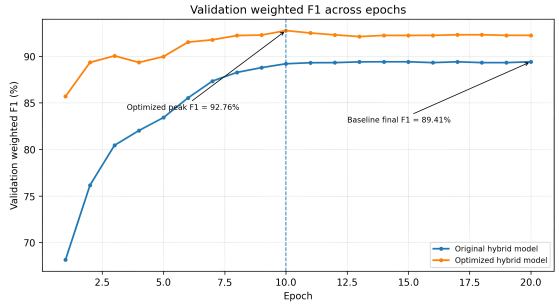


Fig. 6. Validation weighted F1 across epochs for the original and optimized hybrid models.

D. Contextual Comparison with Reported FABSA Baselines

In order to contextualize the proposed model, Table IV compares its performance with baseline and transformer-based

models reported on the FABSA benchmark by Kontonatsios et al. [19]. The proposed BERT–CNN–BiLSTM model achieves precision, recall, and F1-score values of 0.840, 0.890, and 0.870, respectively, which indicate competitive performance in this contextual comparison.

TABLE IV. CONTEXTUAL COMPARISON WITH REPORTED FABSA MODELS.

Model	Precision	Recall	F1-score
LogReg-BoW	0.784	0.491	0.604
GRU-GNN	0.702	0.667	0.684
BERT-single-base	0.785	0.745	0.765
BERT-single-large	0.785	0.792	0.788
BERT-PT	0.785	0.792	0.788
GAS	0.785	0.779	0.782
RoBERTa-single-base	0.781	0.787	0.784
RoBERTa-single-large	0.792	0.816	0.804
RoBERTa-pair-base	0.796	0.764	0.779
RoBERTa-pair-large	0.807	0.792	0.800
DeBERTa-single-base	0.777	0.787	0.782
DeBERTa-single-large	0.791	0.820	0.805
DeBERTa-pair-base	0.793	0.804	0.798
DeBERTa-pair-large	0.806	0.812	0.809
BERT-CNN-BiLSTM Hybrid Model	**0.840**	**0.890**	**0.870**

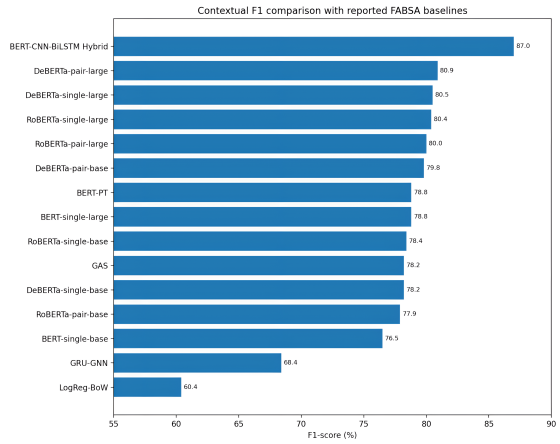


Fig. 7. F1-score comparison between the proposed model and reported FABSA benchmark models.

The proposed model obtains an F1-score of 0.870 in this contextual comparison, while the highest reported baseline F1-score in Table IV is 0.809 for DeBERTa-pair-large. This suggests that the hybrid architecture is promising for aspect-expanded sentiment classification. However, since the baseline models were not reimplemented under the same preprocessing, training, and evaluation conditions, this result should not be interpreted as a definitive ranking. A strict leaderboard comparison would require all models to be evaluated under identical experimental settings (see Fig. 7).

E. Class-Wise Validation Analysis

The optimized model proves to be most effective for the positive class, exhibits consistency for the negative class, and reaches its highest percentage gain for the neutral class. The values of precision, recall, F1-score, and support of the optimized model on the validation set are shown in Table V.

TABLE V. CLASS-WISE VALIDATION PERFORMANCE OF THE OPTIMIZED HYBRID MODEL

Class	Precision	Recall	F1-score	Support
Negative	0.8863	0.8675	0.8768	566
Neutral	0.6914	0.8485	0.7619	66
Positive	0.9534	0.9511	0.9522	1226

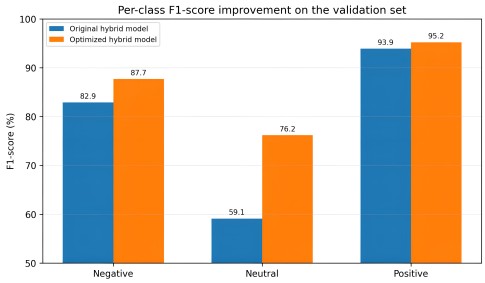
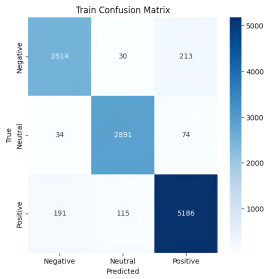
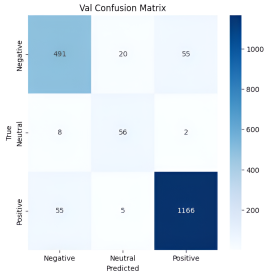


Fig. 8. Per-class F1-score comparison between the original and optimized hybrid models.

As seen in Fig. 8, the highest improvement is observed for the neutral category, with its F1 score growing from 59.1% to 76.2%, which means the absolute improvement reaches 17.1%. For the negative class, the F1 score goes up from 82.9% to 87.7%, and for the positive one, the F1 score jumps from 93.9% to 95.2%. Therefore, this finding confirms the conclusions that can be drawn based on the macro F1 score results.



(a) Training confusion matrix



(b) Validation confusion matrix

Fig. 9. Training and validation confusion matrices of the optimized BERT-CNN-BiLSTM model.

The confusion matrices provided in Fig. 9 show that the optimized model achieves very good class separation on both the train and validation splits, with observations forming a

cluster along the main diagonal of the confusion matrix. On the train data, the number of correctly classified examples for the negative, neutral, and positive classes is 2,514, 2,891, and 5,186 respectively, which shows that the model can learn discriminative representations for all three classes. On the validation split, there are 491 correct predictions for the negative class out of 566, 56 for the neutral class out of 66, and 1,166 for the positive class out of 1,226. The main misclassification sources occur in the classification of positive and negative sentiments, which results in 55 examples being misclassified from one of these two classes to another. While the neutral class is predicted quite accurately, a small portion of negative and positive examples is incorrectly classified as belonging to the neutral class.

F. Cross-Dataset Validation

Cross-dataset validation was used to evaluate generalization. Optimized BERT-CNN-BiLSTM model trained using FABSA dataset was tested on SemEval-2014 Restaurant data without any further training on the external data. SemEval labels were converted for sentiment classification into the same three-class label set used in FABSA dataset: negative, neutral, and positive classes. The SemEval-2014 Restaurant split used in this evaluation consists of 3,602 instances, of which 2,164 were positive, 805 were negative, and 633 were neutral (see Table VI).

TABLE VI. CROSS-DATASET VALIDATION SUMMARY

Evaluation data	Setting	Samples	Accuracy	Weighted F1	Macro F1
FABSA test set	In-domain held-out test	2,812	0.9282	0.9286	0.88
SemEval-2014 Restaurant	Zero-shot cross-dataset	3,602	0.7160	0.6547	0.51

The SemEval-2014 zero-shot evaluation achieves an accuracy of 0.7160 and weighted F1 score of 0.6547. Comparing this to the in-domain FABSA weighted F1 score of 0.9286, the external testing achieves an absolute reduction in F1 score by about 0.2740. This can be expected because of the existence of the domain shift problem where the model has been trained using FABSA dataset reviews but tested using SemEval review without being externally fine-tuned. Despite this domain shift, there still remains considerable sentiment transfer abilities in detecting polarity sentiment.

TABLE VII. CLASS-WISE SEMEVAL-2014 ZERO-SHOT VALIDATION REPORT.

Class	Precision	Recall	F1-score	Support
Negative	0.58	0.73	0.65	805
Neutral	0.42	0.02	0.04	633
Positive	0.77	0.91	0.84	2,164

The class-wise results obtained on SemEval-2014 show that transferability is highest for the positive class (F1 $= 0.84$) and moderate for the negative class (F1 $= 0.65$). However, the neutral class exhibits a near-total performance collapse, with an F1-score of only 0.04, which requires further analysis. This result can be explained by annotation-level semantic divergence between the two datasets. In FABSA, neutral sentiment is defined at the aspect level, where a reviewer may be neutral toward a specific product feature while expressing strong sentiment toward other aspects. In SemEval-2014, neutral polarity is often assigned at the sentence level and may represent factual or descriptive statements without explicit sentiment expression. As a result, the model trained on FABSA-style neutral instances encounters a qualitatively different neutral category during zero-shot transfer. The precision–recall pattern in Table VII further supports this interpretation: neutral precision is 0.42, but recall is only 0.02, indicating that the model rarely assigns the neutral label. Although the class-weighting scheme used during training was intended to improve neutral-class recall within the FABSA domain, the model appears to default to polar predictions when exposed to the more domain-specific vocabulary and sentence-level annotations of SemEval-2014.

A class-prior mismatch may also contribute to this behavior. Neutral instances represent only 17.6% of the SemEval-2014 Restaurant test split, with 633 neutral samples out of 3,602 total instances, whereas the FABSA training distribution was altered through upsampling to improve minority-class representation. This distributional difference may reinforce the model's tendency to avoid neutral predictions under external-domain inference. Taken together, annotation divergence, learned polar decision bias, and class-prior mismatch explain the observed neutral F1-score of 0.04. This finding does not invalidate the overall cross-dataset result, as the model still achieves 71.6% accuracy and a weighted F1-score of 0.65, but it shows that neutral sentiment representations are highly domain-sensitive. Future work should address this issue through aspect-conditioned input formatting, where the aspect term is explicitly included during tokenization, and through domain-adaptive fine-tuning on target-domain neutral examples to recalibrate the decision boundary for this class.

G. Result Interpretation

The experimental results indicate that the proposed hybrid configuration based on the combination of BERT, CNN, and BiLSTM architectures is effective for sentiment classification in the aspect-expanded FABSA review data set. The optimized configuration provided 0.9220 validation accuracy, 0.9225 weighted F1-score, and 0.8636 macro F1-score which are better than original metrics reported for the hybrid architecture. The results of class-wise prediction confirmed the effectiveness of the optimized model since F1-scores were equal to 0.9522, 0.8768, and 0.7619 for positive, negative, and neutral sentiments correspondingly. Even if the neutral sentiment was still the most difficult category to classify, the improved F1-score indicated that the optimization improved the performance and helped to mitigate the effects of the class imbalance in minority sentiments. As for the confusion matrices, the main diagonal showed that the class separation remained stable while the other errors concerned sentiment classes that were contextually ambiguous.

The results of ablation study have confirmed that using the full hybrid configuration contributed to achieving high F1-score. While BERT used alone produced an accurate result

of 0.9167, adding one more layer of either CNN (0.9196) or BiLSTM (0.9182) improved the performance. Finally, the full BERT-CNN-BiLSTM configuration obtained the best result of 0.9286 suggesting that the CNN and BiLSTM layers complemented each other providing local sentiment cues and bidirectional dependencies. The contextual comparison with reported FABSA baselines suggests that the proposed model is competitive, although the comparison cannot be treated as a strict leaderboard result because the baseline models were not reproduced under identical experimental conditions. However, in cross-dataset validation on the SemEval-2014 Restaurant data set, the model showed 0.7160 validation accuracy and 0.6547 weighted F1-score indicating weaker cross-dataset generalization ability. Overall, the results indicate that the proposed method was applicable for conducting context-aware sentiment analysis of e-commerce review data sets.

IV. Discussion

The results of the experiment show that the proposed BERT-CNN-BiLSTM architecture can effectively be used in sentiment classification of FABSA review cases. Specifically, the optimized version of the architecture proved to possess high validation scores of 0.9220 in terms of accuracy, 0.9225 in terms of weighted F1-score and 0.8636 in terms of macro F1-score outperforming the initial hybrid setup in all evaluated metrics. This is confirmed by the class-wise results of the validation, according to which the positive class received the highest F1 score while being followed by the negative and neutral classes. While neutral sentiment appears to be the most difficult to classify among other sentiments, its significant increase in F1-score from 0.591 to 0.7619 shows that optimized training has reduced the impact of class imbalance problem and increased the model's ability to recognize minority sentiment cases.

The ablation experiment supports the efficiency of the full-hybrid approach. In particular, BERT model provided high F1-score already, but the incorporation of either CNN or BiLSTM yielded additional improvements. Finally, BERT-CNN-BiLSTM produced the highest F1-score of 0.9286 suggesting that the addition of both CNN and BiLSTM branches provides additional advantage. The layers of CNN are helpful for recognizing phrase-level sentiment cues, while BiLSTM layers can model dependencies between the parts of the review. Therefore, the proposed architecture can be regarded as efficient both by general and by component-based assessment proving that contextual, local, and sequential models jointly enable more precise sentiment analysis.

Finally, the comparison with existing FABSA baseline models and additional testing on SemEval datasets demonstrate the pros and cons of the proposed model. It demonstrates competitive performance in FABSA review cases comparing with other baseline models of the framework; however, this comparison has to be made under contextual conditions since the leaderboard claim cannot be made due to potential differences in evaluations. However, zero-shot SemEval-2014 Restaurant testing revealed that it has low performance when compared with in-domain tests, which may be caused by the nature of domains. Still, the model retains the ability to correctly recognize polarity for positive and negative sentiments.

V. Limitations

There are several limitations in this study that need to be highlighted. First, although the preprocessing pipeline expands the review data using aspect–sentiment pairs, the aspect term itself is not explicitly appended to the review text during tokenization. Therefore, the current system should be understood as aspect-expanded sentiment classification rather than a fully aspect-conditioned sentiment model. Second, class imbalance remains an important challenge, particularly for the neutral class. Although stratified oversampling and class-weighted loss improve minority-class recognition, the neutral class still shows weaker performance than the positive and negative classes, especially under cross-dataset evaluation. Third, the model is evaluated primarily on the FABSA dataset, with additional zero-shot testing on SemEval-2014 Restaurant data. Broader validation across multiple product categories, domains, languages, and review platforms would be required to confirm the robustness and generalizability of the proposed approach. Finally, the study does not evaluate computational efficiency, inference time, model size, or deployment feasibility, which should be considered before applying the model in real-world e-commerce systems.

VI. Conclusion

This study proposed a hybrid BERT–CNN–BiLSTM model for context-aware sentiment classification of aspect-expanded e-commerce review data. The model combines BERT-based contextual representation learning, CNN-based local feature extraction, and BiLSTM-based bidirectional sequence modeling to capture both semantic and sequential patterns in review text.

Experimental results on the FABSA dataset show that the optimized hybrid model achieves 92.20% validation accuracy, 92.25% weighted F1-score, and 86.36% macro F1-score. The class-wise results further indicate improvement across all sentiment categories, with the strongest gain observed for the neutral class. This suggests that the optimized configuration improves class-balanced sentiment recognition rather than only improving performance on the dominant class.

The architecture-level ablation study confirms that the full BERT–CNN–BiLSTM model performs better than reduced variants using only BERT, BERT–CNN, or BERT–BiLSTM. In addition, the zero-shot evaluation on the SemEval-2014 Restaurant dataset shows that the model retains some ability to recognize sentiment polarity under domain shift, although neutral sentiment remains difficult to transfer across domains. Overall, the findings demonstrate that combining transformer-based contextual encoding with convolutional and recurrent components is an effective approach for sentiment classification in aspect-expanded e-commerce review data.

VII. Future Scope

Future research may extend the proposed sentiment classification framework by incorporating review helpfulness prediction as a separate downstream task or as part of a multi-task learning architecture. Such an extension would require datasets containing helpfulness labels or helpfulness votes, together with appropriate classification or regression-based evaluation metrics [21], [7], [24]. This would make it possible

to examine whether context-aware sentiment representations can also support the estimation of review usefulness.

Further improvement may also be achieved by making the model more explicitly aspect-aware. Although the present study expands review instances using aspect–sentiment pairs, the aspect term is not directly included in the tokenized input sequence. Future work may address this by structuring the input as a review–aspect pair, enabling the model to learn sentiment specifically toward the target aspect. This direction may be especially useful for reviews that contain mixed opinions about different product features [19], [6].

Future studies should also consider broader evaluation settings, including full test-set analysis, cross-validation, cross-domain evaluation, and comparisons across multiple benchmark datasets. Since the zero-shot evaluation on SemEval-2014 Restaurant data shows weaker transferability for the neutral class, additional experiments are needed to better understand the effects of domain shift, annotation differences, and neutral sentiment ambiguity [20]. Class imbalance should also be addressed more thoroughly using methods such as class-balanced loss, focal loss, aspect-level sampling, or targeted data augmentation for underrepresented sentiment classes [13], [20]. These strategies may further improve minority-class recognition, particularly for neutral sentiment. From a practical perspective, future work should also investigate inference speed, model size reduction, calibration, interpretability, and explainability so that the proposed approach can be adapted for real-world e-commerce review analysis.

VIII. Acknowledgment

We would like to express our sincere gratitude to our supervisor, Sifat Rahman Ahona, for the invaluable guidance, continuous support, and insightful feedback provided throughout the course of this research. We also extend our deepest appreciation to the Department of Computer Science and Engineering at the American International University-Bangladesh (AIUB) for providing the necessary facilities and fostering a conducive environment for our work.

References

[1] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of BERT-BiGRU-Softmax," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7819–7837, 2020.

[2] O. Bellar, A. Baina, and M. Ballafkih, "Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers," *Mathematics*, vol. 12, no. 15, p. 2403, 2024.

[3] A. A. Abohany, A. Shora *et al.*, "A hybrid deep learning approach for enhanced sentiment classification and consistency analysis in customer reviews," *Mathematics*, vol. 12, no. 23, p. 3856, 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/23/3856

[4] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Systems with Applications*, vol. 41, p. 3041–3046, 05 2014.

[5] J. Du, J. Rong, S. Michalska, H. Wang, and Y. Zhang, "Feature selection for helpfulness prediction of online product reviews: An empirical study," *PLOS ONE*, vol. 14, no. 12, p. e0226902, 2019.

[6] Y. Zhang, X. Li, and Y. Zhou, "A survey on online reviews in e-commerce: Classification, summarization, and sentiment analysis," *IEEE Access*, vol. 11, pp. 110 829–110 845, 2023.

[7] X. Li, Q. Li, and J. Kim, "A review helpfulness modeling mechanism for online e-commerce: Multi-channel cnn end-to-end approach," *Applied Artificial Intelligence*, vol. 37, no. 1, 2023.

[8] Q. Tul, M. Ali, A. Riaz, A. Noureen, M. Kamranz, B. Hayat, and A. Rehman, "Sentiment analysis using deep learning techniques: A review," *International Journal of Advanced Computer Science and Applications*, vol. 8, 01 2017.

[9] A. M. van der Veen and E. Bleich, "The advantages of lexicon-based sentiment analysis in an age of machine learning," *PLoS ONE*, vol. 20, no. 1, p. e0313092, 2025.

[10] A. A. Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha, and Z. Aung, "Challenges and opportunities of text-based emotion detection: A survey," *IEEE Access*, vol. 12, 2024.

[11] M. M. Taye, R. Abulail, B. Al-Ifan, and F. Alsuhimat, "Enhanced sentiment classification through ontology-based sentiment analysis with bert," *Journal of Internet Services and Information Security (JISIS)*, vol. 15, no. 1, pp. 236–256, 2025.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

[14] J. Shen, "Sentiment analysis of social media texts for aquatic product listed companies based on the BERT-CNN-BiLSTM-Att hybrid model," *Advances in Education, Humanities and Social Science Research*, vol. 13, pp. 182–185, 2025.

[15] L. Khan, A. Qazi, H.-T. Chang, M. Alhajlah, and A. Mahmood, "Empowering Urdu sentiment analysis: An attention-based stacked CNN-BiLSTM deep neural network with multilingual BERT," *Complex & Intelligent Systems*, 2024.

[16] A. Ba Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with turkish universities using twitter data," *Decision Analytics Journal*, vol. 11, p. 100473, 04 2024.

[17] R. Ahamad and D. Mishra, "Exploring sentiment analysis in handwritten and e-text documents using advanced machine learning techniques: a novel approach," *Journal of Big Data*, vol. 12, 01 2025.

[18] A. Angdresey, L. Sitanayah, and I. L. H. Tangka, "Sentiment analysis for political debates on youtube comments using bert labeling, random oversampling, and multinomial naive bayes," *Journal of Computing Theories and Applications*, vol. 2, no. 3, pp. 343–354, 2025.

[19] G. Kontonatsios, J. Clive, G. Harrison, T. Metcalfe, P. Sliwiak, H. Tahir, and A. Ghose, "Fabsa: An aspect-based sentiment analysis dataset of user reviews," *Neurocomputing*, vol. 562, p. 126867, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223009906

[20] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[21] H. Zureigat, O. Al-Qudah, and B. Alhijawi, "Predicting the helpfulness of online customer reviews," in *2023 International Conference on Information Technology (ICIT)*, Aug. 2023, pp. 686–689.

[22] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 08 2014.

[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 01 2018, pp. 328–339.

[24] S. Kim, S. Park, X. Li, and J. Kim, "A deep learning-based review helpfulness prediction system for online beauty products," *IEEE Access*, vol. PP, pp. 1–1, 01 2025.

Advanced and Classical Selection Methods in Genetic Algorithms: A Comprehensive Comparative Analysis

Husam S. Mashaqbeh[1], Putra Sumari[2]*, Hamza A. Mashagba[3], Mohammed Hashem Almourish[4], Azlan Abd Aziz[5], Lara A. Al-Mashagba[6], Wael Waheed Alqassas[7]
Department of Computer Science-Faculty of Information Technology, Isra University, Amman, Jordan[1,4]
School of Computer Science, Universiti Sains Malaysia, 11800 Penang, Malaysia[1,2,6]
Faculty of Engineering and Technology, Multimedia University Melaka, Centre for Wireless Technology (CWT), Malaysia[3,5]
Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen[4]
Computer Science Department, Al al-Bayt University, Mafraq, Jordan[7]

***Abstract*—Selection mechanisms critically influence the convergence behavior and solution quality of Genetic Algorithms (GAs). This study presents a rigorous empirical comparison of six selection methods: three classical methods—Random Selection, Roulette Wheel Selection (RWS), and Tournament Selection (TS)—and three adaptive methods: Fitness-Distance Balance (FDB), Dynamic FDB (dFDB), and Functional Weight-based Selection (FW). Experiments were conducted across 23 classical benchmark functions (F1–F23) and 10 CEC2019 functions (cec01–cec10), with each configuration executed 30 times using consistent GA parameters. Performance was assessed using Best, Mean, Median, and Standard Deviation, with statistical significance determined by the Wilcoxon rank-sum test ($\alpha = 0.05$). The results reveal that TS consistently achieved the best or statistically equivalent performance in 30 out of 33 functions, outperforming both classical and adaptive alternatives. Notably, RWS showed surprising competitiveness, outperforming adaptive methods such as FDB and dFDB in several scenarios. While dFDB and FW improved over static FDB, they failed to consistently outperform TS. These findings confirm TS as a robust default choice for diverse optimization landscapes and provide new empirical evidence regarding the limited practical advantage of current adaptive strategies within GAs. This study contributes the first controlled GA-based evaluation of adaptive selection mechanisms on both classical and CEC2019 benchmarks, offering insights for practitioners designing efficient evolutionary systems. Limitations related to fixed GA settings, function diversity, and adaptive method complexity are acknowledged, and future work is suggested to explore hybrid and problem-aware selection strategies.**

***Keywords*—Genetic algorithm; selection mechanisms; FDB; tournament; roulette wheel; evolutionary computation; optimization**

I. Introduction

Genetic Algorithms (GAs) represent a foundational technique in evolutionary computation, widely applied for solving complex optimization tasks by emulating the processes of natural selection and evolution [1]. A key determinant of GA performance is the selection operator, which mediates the trade-off between exploration of the solution space and exploitation of high-quality individuals [2]. Over time, a diverse range of selection methods has been developed—from traditional approaches like Roulette Wheel and Tournament Selection [2], [3], to more advanced, adaptive techniques such as the FDB family, including FDB, dFDB, and FW methods [4], [5]. These adaptive methods dynamically adjust selection pressure based on population distribution and fitness, potentially improving convergence and mitigating premature stagnation.

While considerable attention has been given to modifying GA components—including crossover, mutation, and population structure—relatively few studies have systematically compared selection strategies across a broad spectrum of problem types. Notably, distance-aware selection mechanisms have gained popularity in other metaheuristics, including LSHADE [6], Whale Optimization Algorithm (WOA) [7], and Stochastic Fractal Search (SFS) [8], demonstrating promising outcomes. However, their effectiveness within the traditional GA framework remains underexplored. Moreover, although GAs have been successfully applied in various fields, from photovoltaic modeling [9] to power system optimization [10], the optimal choice of selection strategy under consistent algorithmic settings is still an open question.

This study addresses this gap by systematically evaluating six selection methods—three classical and three adaptive—within a canonical GA setup. The goal is to empirically assess their comparative performance across diverse benchmarks and derive practical recommendations for researchers and practitioners. The primary contributions of this work are:

- Comprehensive Comparison of Selection Strategies: This work presents one of the first integrated empirical comparisons between classical GA selection methods (Random, Roulette Wheel Selection, Tournament Selection) and adaptive FDB-based methods (FDB, dFDB, FW), under a unified experimental setup.
- Diverse and Challenging Benchmarking: The evaluation spans 23 classical benchmark functions and 10 CEC2019 functions, encompassing unimodal, multi-

*Corresponding author.

modal, composite, and high-dimensional landscapes.

- Robust Statistical Validation: In addition to standard performance metrics (Best, Mean, Median, Standard Deviation), the study employs the Wilcoxon rank-sum test to rigorously assess the statistical significance of observed differences between selection strategies.
- In-Depth Literature Integration: The study synthesizes and contextualizes prior research on both classical and modern selection approaches, clarifying existing gaps.
- Actionable Insights for Practitioners: By holding all other GA components constant and varying only the selection mechanism, the study isolates its effect on performance, demonstrating that selection alone can significantly impact results.

The remainder of this study is organized as follows: Section II reviews classical and advanced selection methods. Section III describes the proposed methodology. Section IV presents experimental results and statistical analysis. Section V discusses key findings and their implications. Section VI concludes the study and outlines future research directions.

II. Literature Review

This section briefly reviews key developments in selection methods, beginning with classical techniques, followed by recent adaptive strategies based on the FDB framework.

A. Classical Selection Methods

Classical selection strategies in GAs primarily include proportional selections such as RWS and ranking-based methods, including TS. While simple and computationally efficient, RWS is highly sensitive to skewed fitness distributions; a small number of very fit individuals can dominate the mating pool, increasing the risk of premature convergence [2]. TS, in contrast, applies controlled selection pressure and has been widely adopted for its robustness, but it does not inherently incorporate population diversity or spatial distribution of solutions. As noted by Shukla et al. [11], while tournament size can modulate pressure, classical methods generally ignore diversity, which can hinder their performance on multi-modal or complex landscapes.

In response to these limitations, enhancements have been proposed. For example, Cuevas et al. [12] introduced a Golden Section Selection inspired by the golden ratio, which outperformed traditional schemes on selected benchmarks by improving solution quality by approximately 15–20%. Likewise, Oladimeji et al. [13] demonstrated performance gains in the Mayfly Algorithm by replacing random selection with RWS.

Table I summarizes key studies on classical selection methods.

TABLE I. Summary of Classical Selection Method Studies

Reference	Focus	Methodology	Results
Goldberg & Deb (1991) [2]	Comparative analysis of selection schemes	Modeling and simulations	Tournament-based methods exert stronger selection pressure
Julstrom (1999)[3]	Rank-based vs. TS equivalence	Analytical probability-based analysis	TS preferred due to simplicity and consistent pressure
Zhang & Kim (2001)[14]	Selection methods for layout optimization	Experiments on machine layout problems	TS and ranking selection outperformed others
Zhong et al. (2005) [15]	RWS vs. TS on benchmark functions	Benchmark tests with fixed settings	TS converged faster than RWS
Mashohor et al. (2005) [16]	Selection for industrial inspection	Evaluation of PCB inspection system	Elitist selection gave the best results
Shukla et al. (2015) [11]	Review of selection techniques	Conceptual and experimental analysis	TS incurred lower overhead for similar results
Yadav et al. (2017) [17]	Multiple selection schemes	Conceptual and partial experiments	TS and elitism yielded faster convergence
Cuevas et al. (2018) [12]	Golden section vs. traditional methods	Benchmark optimization	Golden-section improved quality by ∼15–20%
Oladimeji et al. (2024)[13]	RWS in Mayfly algorithm	Benchmarked on face/iris datasets	RWS improved accuracy and reduced error

B. Advanced Selection Methods (FDB Family)

To address the limitations of classical selection strategies, the FDB family introduces a mechanism for maintaining population diversity by jointly considering an individual's fitness and its spatial distance from the best-performing solution. The original FDB method, introduced by Kahraman et al. in 2019 [4], assigns each individual a composite score that combines normalized fitness and distance values. The general FDB scoring equation is:

$$\text{Score}_i = \delta \cdot \text{normF}_i + (1 - \delta) \cdot \text{normD}_i \tag{1}$$

where, normF and normD are the normalized fitness and distance scores, respectively, and $\delta \in [0, 1]$ is a weight controlling the trade-off between exploitation and exploration.

Multiple variants of FDB have since been developed. One notable extension is Fitness-Distance Balance with Functional Weights (FW), proposed by Tao et al. in 2021 [5]. Rather than using a fixed weighting parameter, FW introduces a stochastic weighting scheme where δ is sampled from a probability distribution (e.g., Gaussian or Cauchy) during each generation.

Another enhancement is Dynamic Fitness-Distance Balance (dFDB), introduced by Kahraman et al. [8], which varies the weight δ over time according to a predefined schedule. Typically, dFDB starts with greater emphasis on diversity (higher distance weight) and gradually shifts toward fitness-based exploitation as the search progresses.

Table II summarizes key contributions utilizing the FDB family of methods.

Research Gap: While classical selection strategies such as RWS and TS have been extensively evaluated on standard benchmark functions, their effectiveness on large-scale or highly complex optimization problems remains insufficiently studied. Conversely, adaptive methods—including FDB, dFDB, and FW—have demonstrated strong potential in modern metaheuristics, but their systematic integration into and evaluation within traditional GAs is still lacking.

TABLE II. SUMMARY OF FDB-BASED SELECTION METHODS

Reference	Focus of Study	Methodology	Key Findings
Kahraman et al. (2019)[4]	Introduction of FDB	Applied in various meta-heuristics (90 benchmarks)	FDB improved diversity and solution quality
Tao et al. (2021)[5]	FW for dynamic balance	Integration into spherical search algorithm	FW outperformed traditional methods
Tan et al. (2022) [18]	Chaotic FDB-based selection	Hybrid wind-driven optimization with FDB	Enhanced accuracy and faster convergence
Battal & Guvenc (2022) [6]	FDB-integrated LSHADE	Energy hub management	Superior performance in energy optimization
Hou et al. (2024) [7]	Adaptive FDB in WOA	Modified WOA with FDB component	Increased accuracy and faster convergence
Kahraman et al. (2024) [19]	dFDB for SFS	Applied to SFS and photovoltaic optimization	dFDB provided superior robustness
Kahraman et al. (2023)[19]	dFDB for GBO	Incorporated dFDB into GBO	Improved exploration and balanced search
Mashaqbeh & Mjlae (2025) [20]	FDB-enhanced Jaya	Cloud task scheduling	Reduced makespan by up to 38.98%

TABLE III. CONFIGURATION AND PARAMETERS USED FOR ALL EXPERIMENTS

Parameter	Value
Population Size	75
Generations	3000
Encoding	Real-valued vector
Benchmark Functions	CEC2019 and Classical (F1–F23)
Dimensions	10 (CEC2019); varies (F1–F23)
Crossover Type	Blend crossover (BLX-0.5)
Crossover Rate	0.7
Mutation Type	Uniform mutation
Mutation Rate	0.3 (per individual)
Runs	30
Statistical Test	Wilcoxon Rank-Sum Test ($\alpha = 0.05$)
Performance Metrics	Best, Mean, Median, Std
Random	Uniformly random
RWS	Proportional to fitness
TS	TS size $= 3$
FDB	Static weight $\delta = 0.5$
dFDB	δ decreases every frequency $= 10$
FW	Gaussian sampling for dynamic δ

This study bridges this gap through a unified and controlled comparison across both traditional and CEC2019 benchmark sets.

III. METHODOLOGY

This study employed a standardized GA implemented in MATLAB to ensure a consistent experimental framework across all comparisons. The only varying component was the selection method, while all other algorithmic elements were held constant. Table III summarizes the GA configuration and selection method parameters used throughout the experiments.

The GA evolved with a population of 75 individuals over 3000 generations. Each individual was encoded as a real-valued vector, with bounds and dimensions determined by the benchmark function. For reproduction, blend crossover (with a rate of 0.7 and gamma in the range $[-0.5, 1.5]$) and uniform mutation (applied with a 0.3 probability per individual) were used.

Six selection strategies were tested:

- Random: No selection pressure; individuals were randomly chosen.
- RWS: Proportional selection based on normalized fitness.
- TS: A tournament of size 3 selects the best among randomly sampled candidates.
- FDB: Applies a composite score based on static weight $\delta = 0.5$ combining normalized fitness and distance to the current best.
- dFDB: The weight δ decreases linearly from 0.6 to 0.0 over the total number of generations to balance exploration and exploitation dynamically.
- FW: Uses a stochastic weight $\delta \sim \mathcal{N}(0.5, 0.1)$ resampled each generation, providing adaptive selection pressure.

The benchmark consisted of two test suites:

- F1–F23: Classical benchmark functions covering unimodal, multimodal, and hybrid categories.
- CEC2019 (cec01–cec10): Designed to challenge optimization strategies through transformed, rotated, and composite problem landscapes, implemented at 10 dimensions.

For each function and selection method pair, 30 independent runs were executed. Performance was evaluated based on the Best, Mean, Median, and Std of the final fitness values. Additionally, the Wilcoxon rank-sum test was applied to assess statistical significance between each method and TS, which served as a robust baseline due to its proven performance in preliminary evaluations.

IV. RESULTS

This section presents a comprehensive evaluation of the six selection methods across two benchmark suites: the classical functions F1–F23 and the CEC2019 test functions (cec01–cec10). The evaluation is based on multiple statistical indicators (Best, Mean, Std, Median) and supported by Wilcoxon rank-sum tests for statistical significance. Convergence curves and boxplots are included to visualize the performance dynamics and distributional patterns.

A. Performance on CEC2019 Benchmark Functions (cec01–cec10)

The CEC2019 benchmark suite includes complex, composition-based, and real-world inspired optimization problems characterized by large search spaces, deceptive landscapes, and challenging global minima. Table IV presents the performance metrics, and Table V reports the Wilcoxon test results against TS.

TABLE IV. RESULTS ON CEC2019 (BEST, MEAN, MEDIAN, STD OVER 30 RUNS)

Fun	Measure	FDB	dFDB	Random	RWS	TS	FW
cec01	Best	48.4E+6	10.3E+6	41.4E+6	923.6E+3	**155.6E+3**	843.1E+3
	Mean	978.7E+6	125.6E+6	1.7E+9	4.2E+6	**1.5E+6**	4.8E+6
	Median	606.6E+6	106.5E+6	534.0E+6	3.1E+6	**1.2E+6**	3.6E+6
	Std	906.8E+6	89.9E+6	2.3E+9	3.6E+6	**1.1E+6**	3.6E+6
cec02	Best	21.3E+0	17.5E+0	49.7E+0	17.3E+0	**17.3E+0**	17.3E+0
	Mean	56.5E+0	17.8E+0	124.7E+0	17.3E+0	**17.3E+0**	17.3E+0
	Median	51.8E+0	17.8E+0	107.9E+0	17.3E+0	**17.3E+0**	17.3E+0
	Std	19.6E+0	240.9E-3	51.2E+0	1.2E-3	**246.0E-6**	1.1E-3
cec03	Best	12.7E+0	12.7E+0	12.7E+0	12.7E+0	**12.7E+0**	12.7E+0
	Mean	12.7E+0	12.7E+0	12.7E+0	12.7E+0	**12.7E+0**	12.7E+0
	Median	12.7E+0	12.7E+0	12.7E+0	12.7E+0	**12.7E+0**	12.7E+0
	Std	582.2E-6	348.9E-6	1.6E-3	154.5E-9	**51.7E-9**	194.3E-9
cec04	Best	2.5E+3	86.7E+0	21.0E+3	11.1E+0	**5.0E+0**	17.1E+0
	Mean	5.7E+3	956.7E+0	59.9E+3	43.7E+0	**46.5E+0**	61.9E+0
	Median	5.6E+3	617.4E+0	57.0E+3	38.6E+0	**44.3E+0**	45.4E+0
	Std	1.9E+3	949.3E+0	32.4E+3	18.6E+0	**24.2E+0**	38.6E+0
cec05	Best	2.1E+0	1.2E+0	3.9E+0	1.0E+0	**1.0E+0**	1.1E+0
	Mean	2.7E+0	2.1E+0	9.9E+0	1.2E+0	**1.0E+0**	1.3E+0
	Median	2.7E+0	2.1E+0	9.4E+0	1.2E+0	**1.0E+0**	1.2E+0
	Std	324.5E-3	384.1E-3	3.7E+0	170.9E-3	**19.3E-3**	172.1E-3
cec06	Best	11.1E+0	7.3E+0	11.7E+0	2.4E+0	**1.8E+0**	2.3E+0
	Mean	14.2E+0	9.4E+0	14.5E+0	3.3E+0	**4.5E+0**	3.4E+0
	Median	14.1E+0	9.5E+0	14.6E+0	3.3E+0	**4.7E+0**	3.2E+0
	Std	1.7E+0	1.1E+0	1.3E+0	625.7E-3	**1.2E+0**	649.3E-3
cec07	Best	1.1E+3	275.0E+0	1.2E+3	20.9E+0	**-61.8E+0**	-272.2E+0
	Mean	1.8E+3	709.1E+0	2.0E+3	241.1E+0	**231.7E+0**	159.4E+0
	Median	1.7E+3	698.9E+0	2.0E+3	231.3E+0	**217.5E+0**	121.2E+0
	Std	422.5E+0	275.0E+0	311.7E+0	151.0E+0	**143.5E+0**	165.8E+0
cec08	Best	6.9E+0	5.0E+0	6.6E+0	4.7E+0	**1.2E+0**	3.6E+0
	Mean	8.2E+0	7.1E+0	8.0E+0	5.6E+0	**4.0E+0**	5.2E+0
	Median	8.2E+0	7.1E+0	8.1E+0	5.7E+0	**4.1E+0**	5.4E+0
	Std	635.2E-3	1.0E+0	454.7E-3	453.5E-3	**1.1E+0**	681.5E-3
cec09	Best	137.3E+0	6.1E+0	1.8E+3	2.6E+0	**2.5E+0**	2.6E+0
	Mean	928.0E+0	193.9E+0	7.2E+3	3.1E+0	**2.9E+0**	3.1E+0
	Median	947.6E+0	159.3E+0	6.0E+3	3.1E+0	**2.9E+0**	3.1E+0
	Std	328.3E+0	191.3E+0	4.3E+3	278.4E-3	**210.9E-3**	240.6E-3
cec10	Best	20.5E+0	20.2E+0	20.6E+0	2.0E+0	**3.4E+0**	20.0E+0
	Mean	20.9E+0	20.4E+0	21.0E+0	19.4E+0	**19.5E+0**	20.0E+0
	Median	20.9E+0	20.5E+0	21.0E+0	20.0E+0	**20.0E+0**	20.0E+0
	Std	219.2E-3	149.6E-3	152.8E-3	3.3E+0	**3.0E+0**	1.3E-3

As with the classical functions, TS maintained its dominance in this suite. It achieved the best or second-best fitness values on nearly all functions, with low variance and high reliability. FW again emerged as the closest competitor, achieving statistically indistinguishable performance on some functions (e.g., cec04, cec06, cec08) with p-values above the 0.05 threshold. RWS showed a more variable pattern: it was competitive on some easier CEC functions (e.g., cec02, cec05), but less reliable on harder ones such as cec01 and cec07.

The adaptive FDB and dFDB methods exhibited lower performance overall. While dFDB improved upon static FDB, especially on functions like cec05 and cec09, both methods lagged significantly behind TS, FW, and RWS. Their higher standard deviations and poorer best values indicate instability and insufficient convergence, particularly under fixed computational budgets.

Convergence curves (Fig. 1) highlight the rapid and stable convergence of TS and FW, in contrast to the delayed and irregular behavior of the adaptive methods. Boxplots (Fig. 2) for cec01–cec10 further illustrate these trends.

B. Performance on Classical Benchmark Functions (F1–F23)

The classical suite includes a variety of unimodal, multimodal, hybrid, and composition functions. Table VI,VII and

TABLE V. WILCOXON RANK-SUM TEST (p-VALUES) FOR CEC2019 VS. TOURNAMENT

Func.	TS vs RWS	TS vs FDB	TS vs dFDB	TS vs Rand.	TS vs FW
cec01	2.00E-06 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	5.60E-07 (S)
cec02	1.09E-10 (S)	3.02E-11 (S)	3.01E-11 (S)	3.02E-11 (S)	5.48E-11 (S)
cec03	7.04E-07 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	8.84E-07 (S)
cec04	8.65E-01 (NS)	3.02E-11 (S)	3.34E-11 (S)	3.02E-11 (S)	2.01E-01 (NS)
cec05	6.12E-10 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	4.98E-11 (S)
cec06	1.34E-05 (S)	3.02E-11 (S)	3.34E-11 (S)	3.02E-11 (S)	2.43E-05 (S)
cec07	8.30E-01 (NS)	3.02E-11 (S)	1.70E-09 (S)	3.02E-11 (S)	3.27E-02 (S)
cec08	1.16E-07 (S)	3.02E-11 (S)	1.61E-10 (S)	3.02E-11 (S)	2.43E-05 (S)
cec09	4.71E-04 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	1.17E-03 (S)
cec10	9.93E-02 (NS)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	2.71E-02 (S)

S = Significant; NS = Not Significant ($\alpha = 0.05$)

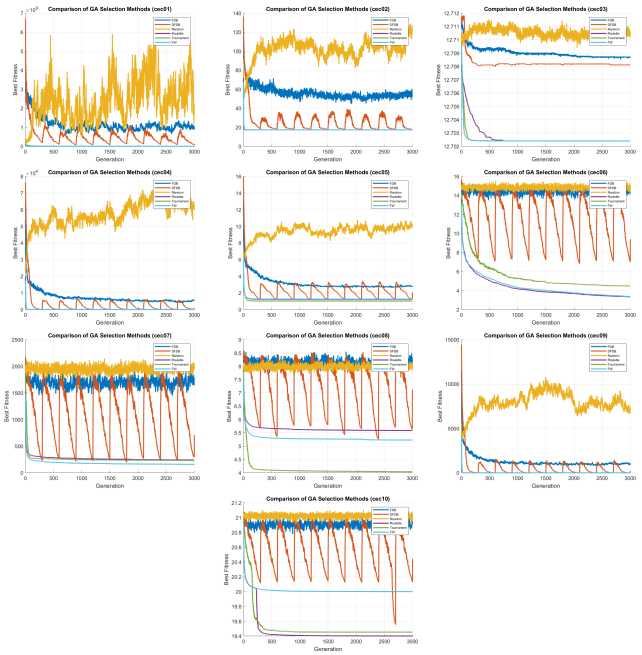


Fig. 1. Convergence curves of selection methods in the optimization of the CEC2019 suite.

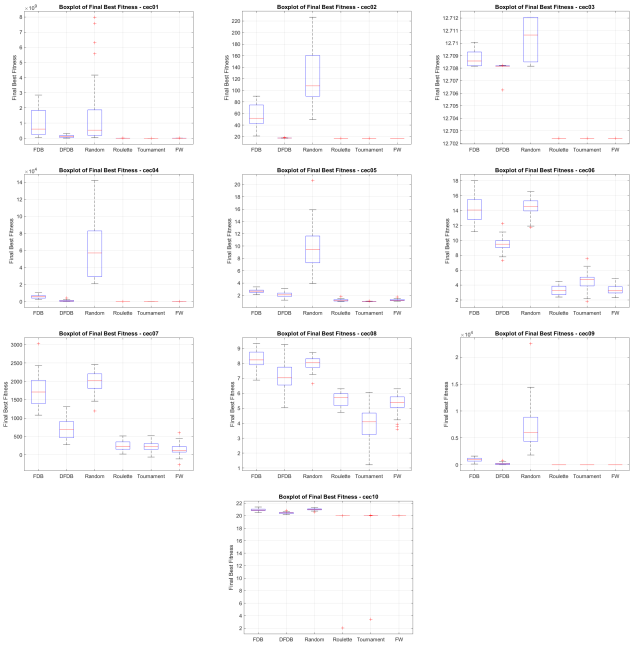


Fig. 2. Boxplot obtained from selection methods on CEC2019 suite.

VIII summarizes the best, mean, and standard deviation of the final fitness values for each function over 30 independent runs.

Across the majority of functions, the TS technique achieved the most consistent performance. This was particularly evident in unimodal functions such as F1 (Sphere) and F6 (Step), where TS achieved optimal solutions in all runs with negligible variance. Notably, FW closely followed TS in most cases, often achieving statistically indistinguishable performance as confirmed by Wilcoxon tests ($p > 0.05$ in many functions such as F8 and F10).

RWS, while simpler and non-adaptive, exhibited mixed performances. It outperformed both adaptive methods (FDB and dFDB) on a significant subset of functions, particularly on F1–F7. The static FDB method consistently yielded the weakest results across both unimodal and multimodal functions, indicating convergence stagnation and poor exploitation. The dynamic variant, dFDB, demonstrated improvements over its static counterpart in some multi-modal problems (e.g., F13, F20), but remained inferior to TS, FW, and RWS in the majority of scenarios.

Statistical analysis (Table VIII) supports these findings. TS significantly outperformed all other methods ($p < 0.05$) on the vast majority of functions. FW was the only method that frequently achieved statistically similar results to TS. Visual inspection of convergence curves (Fig. 3) and boxplots (Fig. 4) corroborates the numerical results.

C. Statistical Summary and General Observations

A cross-suite statistical perspective reveals the following insights:

TABLE VI. RESULTS ON F1–F12 (BEST, MEAN, MEDIAN, STD)

Fun	Measure	FDB	dFDB	Random	RWS	TS	FW
F1	Best	3.86E+03	3.65E+02	2.98E+04	1.22E+00	**3.56E-04**	1.36E+00
	Mean	7.37E+03	1.25E+03	5.27E+04	2.56E+00	**5.27E-03**	2.73E+00
	Median	6.82E+03	1.23E+03	5.03E+04	2.65E+00	**5.08E-03**	2.73E+00
	Std	1.99E+03	6.12E+02	9.83E+03	6.84E-01	**3.56E-03**	7.11E-01
F2	Best	4.19E+03	4.43E+00	4.25E+03	4.12E-01	**2.23E-03**	5.01E-01
	Mean	2.34E+16	9.53E+00	5.17E+12	5.86E-01	**1.30E-02**	6.29E-01
	Median	5.35E+10	9.37E+00	1.06E+09	5.84E-01	**1.31E-02**	6.20E-01
	Std	1.21E+17	2.04E+00	2.03E+13	8.16E-02	**7.62E-03**	7.47E-02

TABLE VII. RESULTS ON F13–F23 (BEST, MEAN, MEDIAN, STD)

Fun	Measure	FDB	dFDB	Random	RWS	TS	FW
F13	Best	4.84E+07	3.90E+04	1.26E+08	4.37E-02	**5.91E-06**	6.22E-02
	Mean	1.53E+08	1.81E+06	6.24E+08	9.02E-02	**3.83E-03**	1.15E-01
	Median	1.36E+08	7.68E+05	6.26E+08	8.71E-02	**2.63E-04**	1.17E-01
	Std	6.69E+07	2.83E+06	2.18E+08	3.10E-02	**5.26E-03**	2.94E-02

TABLE VIII. WILCOXON RANK-SUM TEST (p-VALUES) FOR F1–F23 VS. TOURNAMENT

Fun	TS vs RWS	TS vs FDB	TS vs dFDB	TS vs Rand.	TS vs FW
F1	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F2	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F3	6.70E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	9.92E-11 (S)
F4	4.08E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F5	3.56E-04 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	7.66E-05 (S)
F6	2.70E-03 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	5.20E-10 (S)
F7	4.71E-04 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	2.62E-03 (S)
F8	4.73E-01 (NS)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	7.51E-01 (NS)
F9	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F10	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F11	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F12	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F13	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)
F14	1.90E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	1.90E-11 (S)
F15	3.18E-04 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	4.94E-05 (S)
F16	7.88E-12 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	7.88E-12 (S)
F17	2.26E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	2.50E-11 (S)
F18	2.99E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	2.99E-11 (S)
F19	2.57E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	2.57E-11 (S)
F20	6.05E-07 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	3.52E-07 (S)
F21	5.96E-05 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	6.75E-05 (S)
F22	4.21E-04 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	7.68E-04 (S)
F23	8.18E-07 (S)	3.02E-11 (S)	3.02E-11 (S)	3.02E-11 (S)	9.03E-08 (S)

S = Significant; NS = Not Significant ($\alpha = 0.05$)

- TS consistently achieved top results across both benchmarks, confirming its robustness and effectiveness in both simple and complex landscapes.
- FW, although relatively simple, showed strong performance. It was statistically equivalent to TS in many cases and could serve as a viable alternative, particularly in problems where slightly more exploration is desirable.
- RWS performed better than expected. It surpassed both adaptive methods in many tests and was statistically competitive with TS on select functions.
- dFDB improved upon the static variant but remained inconsistent. It showed promise in maintaining diversity but lacked the convergence efficiency of TS and FW.
- Static FDB and Random Selection were the least effective. They struggled with both exploitation and stability, particularly in high-dimensional and multimodal functions.

Fig. 3. Convergence curves of selection methods in the optimization of functions F1 to F23.

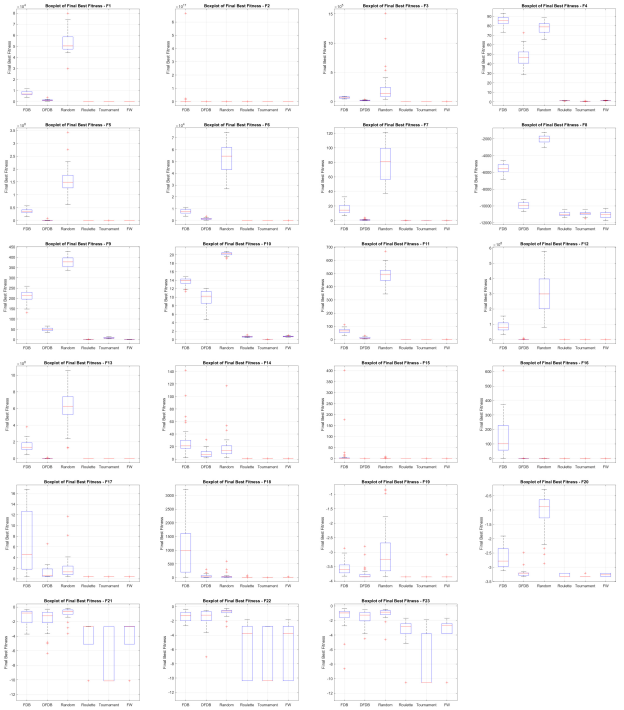


Fig. 4. Boxplot diagram obtained from selection methods on functions F1 to F23.

V. Discussion

A. Robustness and Efficiency of Tournament Selection

The TS consistently emerged as the most effective method across both benchmark suites. These results can be attributed to TS's balanced exploitation-exploration tradeoff, particularly with the moderate tournament size ($k = 3$) used in this study. The deterministic pressure of TS ensures that fitter individuals are consistently favored, facilitating rapid convergence in unimodal and smooth landscapes (e.g., F1, F6, cec02), while still preserving diversity enough to maintain effectiveness in multimodal problems.

B. Competitiveness of Functional Weight-Based Selection

FW Selection demonstrated competitive performance—often statistically indistinguishable from TS on both classical and CEC functions. The method introduces probabilistic selection that is biased by fitness scores but retains stochasticity to avoid premature convergence. This allows FW to maintain a broader exploration radius compared to TS, which can be advantageous in deceptive or rugged landscapes.

C. Re-Evaluating Roulette Wheel Selection

RWS, despite its simplicity and non-adaptive nature, demonstrated surprisingly strong performance on several classical functions, particularly unimodal and moderately multimodal ones (e.g., F1–F5, F10). This challenges the common assumption that adaptive strategies always outperform classical ones. However, RWS also exhibited more variability on highly multimodal or deceptive landscapes.

D. Random Selection as a Baseline

As expected, Random Selection performed poorly across all benchmarks, reinforcing the necessity of selection pressure in evolutionary search [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. Interestingly, Random outperformed static FDB in a few instances (e.g., F18, F19), highlighting that without appropriate tuning, even diversity-focused selection may lead to inefficient search.

E. Limitations of FDB and dFDB

The FDB selection techniques, particularly the static variant, were consistently outperformed by other methods [31], [32], [33], [34], [35], [36], [37], [38], [39]. The static FDB struggled to converge effectively, often yielding large standard deviations and poor mean fitness values. This was evident in functions such as F2, F5, F12, and cec01. The likely cause is the fixed weighting between fitness and diversity: while diversity helps in early search stages, excessive emphasis on it can hinder convergence in later stages. The dFDB showed partial improvement but still lagged behind in most scenarios.

F. Implications for Algorithm Design

The comparative results suggest that classical methods such as TS and RWS, despite their simplicity, remain highly effective across diverse problem domains [40], [41], [42]. The findings also imply that hybrid strategies, such as switching from exploration-heavy to exploitation-heavy selection over time, may outperform static or single-strategy methods.

VI. Conclusion

This study has presented a rigorous comparative evaluation of six selection strategies in GAs, encompassing classical strategies (TS and RWS), adaptive mechanisms (FW, FDB, and dFDB), and a baseline random strategy. Across 33 benchmark functions—including the classical F1–F23 suite and the CEC2019 set—the TS method demonstrated superior robustness, achieving the best or statistically comparable results in the majority of cases. FW method emerged as a strong alternative, often yielding statistically indistinguishable performance with the added benefit of enhanced exploration through adaptive weight control.

Interestingly, RWS, despite its non-adaptive nature, outperformed the adaptive FDB and dFDB methods on a substantial number of functions, highlighting its retained competitiveness in structured problem landscapes. In contrast, the FDB-based strategies underperformed in both convergence and consistency, suggesting the need for better tuning or hybridization. Random Selection illustrated the essential role of selection pressure in evolutionary search.

A. Limitations of the Study

Despite its comprehensive scope, all evaluations were conducted under fixed population sizes and generations, which may constrain generalizability. Furthermore, the benchmark functions used do not fully reflect the stochasticity and complexity of real-world applications. The implementations of FDB and dFDB used fixed parameter schedules, which might not be optimal across all problem types.

B. Future Work

The findings suggest several promising directions: development of hybrid selection strategies combining the strengths of TS with diversity mechanisms; integration of adaptive scheduling or self-regulating parameters; and exploring the interplay between selection methods and other GA components such as crossover operators and mutation schemes.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Acknowledgment

The authors want to acknowledge the financial support under the grant from Multimedia University (MMU) Postdoc Fellowship Fund under the grant number MMUI/250023.

References

[1] A. Brindle, "Genetic algorithms for function optimization," Ph.D. dissertation, University of Alberta, 1981.

[2] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," in *Foundations of Genetic Algorithms*. Morgan Kaufmann, 1991, pp. 69–93.

[3] B. A. Julstrom, "It's all the same to me: Revisiting rank-based probabilities and tournament selection," in *Proceedings of the 1999 Congress on Evolutionary Computation (CEC 1999)*, vol. 2. IEEE, 1999, pp. 1501–1507.

[4] H. T. Kahraman, S. Aras, and E. Gedikli, "Fitness-distance balance (FDB): A new selection method for meta-heuristic search algorithms," *Knowledge-Based Systems*, vol. 187, p. 105169, 2020.

[5] S. Tao, Y. Todo, R.-L. Wang, K. Wang, and S. Gao, "Fitness-distance balance with functional weights: A new selection method for evolutionary algorithms," *IEICE Transactions on Information and Systems*, vol. E104-D, no. 10, pp. 1789–1792, 2021.

[6] O. Battal and U. Guvenc, "Fitness-distance balance based LSHADE algorithm for energy hub management," *IEEE Access*, vol. 10, pp. 70 010–70 023, 2022.

[7] C. Hou, Z. Lei, B. Zhang, and S. Gao, "An improved whale optimization algorithm with adaptive fitness-distance balance," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 19, no. 4, pp. 560–568, 2024.

[8] H. T. Kahraman, M. H. Hassan, M. Katı, M. Tostado-Véliz, and S. Kamel, "A dynamic fitness-distance balance based stochastic fractal search for solving global optimization and determining accurate modeling of photovoltaic models," *Soft Computing*, 2024.

[9] Y. He, K. Wang, and S. Gao, "Robust parameter identification based on nature-inspired optimization for accurate photovoltaic modelling under different operating conditions," *Journal of Renewable and Sustainable Energy*, vol. 16, no. 1, 2024.

[10] N. H. Khan, Y. Wang, R. Jamal, and Z. El-Barbary, "Solving optimal power flow frameworks using modified artificial rabbit optimizer," *International Journal of Energy Research*, 2024.

[11] A. Shukla, H. M. Pandey, and D. Mehrotra, "Comparative review of selection techniques in genetic algorithm," in *Proceedings of the International Conference on Advances in Computer Engineering and Applications (ICACEA)*. IEEE, 2015, pp. 135–140.

[12] E. Cuevas, L. Enríquez, D. Zaldívar, and M. Pérez-Cisneros, "A selection method for evolutionary algorithms based on the Golden Section," *Expert Systems with Applications*, vol. 106, pp. 183–196, 2018.

[13] A. Oladimeji *et al.*, "Effect of roulette wheel selection method on mayfly algorithm," *Journal of AI & Algorithms*, vol. 22, pp. 1–8, 2024.

[14] B. T. Zhang and J. J. Kim, "Comparison of selection methods for evolutionary optimization," *Machine Layout: Evolutionary Optimization on the Internet*, vol. 2, no. 1, pp. 55–70, 2001.

[15] J. Zhong, X. Hu, M. Gu, and J. Zhang, "Comparison of performance between different selection strategies on simple genetic algorithms," in *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA)*, 2005, pp. 982–987.

[16] S. Mashohor, J. R. Evans, and T. Arslan, "Elitist selection schemes for genetic algorithm based printed circuit board inspection system," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2005)*, 2005, pp. 1835–1842.

[17] S. L. Yadav, A. Sohal, and S. Lata, "Comparative study of different selection techniques in genetic algorithm," *International Journal of Engineering, Science and Mathematics*, vol. 6, no. 3, pp. 174–180, 2017.

[18] Y. Tan, Y. Xue, and L. Jiao, "Chaotic wind driven optimization with fitness distance balance-based selection strategy for global optimization," *Cognitive Computation*, vol. 14, no. 1, pp. 327–341, 2022.

[19] H. T. Kahraman, M. H. Hassan, M. Katı, M. Tostado-Véliz, and S. Kamel, "Improved gradient-based optimizer with dynamic fitness distance balance for global optimization problems," *Soft Computing*, 2023.

[20] H. S. Mashaqbeh and S. A. Mjlae, "FL-Jaya scheduler: an optimized Jaya algorithm for effective independent task scheduling in cloud systems," *Computing*, vol. 107, p. 129, 2025.

[21] Y. Z. Rawash, B. Al-Naami, A. Alfraihat, and H. A. Owida, "Advanced low-pass filters for signal processing: A comparative study on gaussian, mittag-leffler, and savitzky-golay filters." *Mathematical Modelling of Engineering Problems*, vol. 11, no. 7, 2024.

[22] F. Alnaimat, H. Owida, A. Al Sharah, M. Alhaj, and M. Hassan, "Silicone and pyrocarbon artificial finger joints," *Applied Bionics and Biomechanics*, vol. 2021, no. 1, p. 5534796, 2021.

[23] G. AlMahadin, H. A. Owida, J. I. Al-Nabulsi, N. Turab, and N. Al Hawamdeh, "Automated detection of kidney masses lesions using a deep learning approach," *Int. J. Artif. Intell. ISSN*, vol. 2252, p. 2863, 2024.

[24] H. A. Owida, H. S. Migdadi, O. S. M. Hemied, N. F. F. Alshdaifat, S. F. A. Abuowaida, and R. S. Alkhawaldeh, "Deep learning algorithms to improve covid-19 classification based on ct images," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2876–2885, 2022.

[25] H. Abu Owida, B. Moh'd, N. Turab, J. Al-Nabulsi, and S. Abuowaida, "The evolution and reliability of machine learning techniques for oncology," *International Journal of Online and Biomedical Engineering (iJOE)*, 2023.

[26] H. Abu Owida, "Developments and clinical applications of biomimetic tissue regeneration using 3d bioprinting technique," *Applied Bionics and Biomechanics*, vol. 2022, no. 1, p. 2260216, 2022.

[27] N. Alshdaifat, H. A. Owida, Z. Mustafa, A. Aburomman, S. Abuowaida, A. Ibrahim, and W. Alsharafat, "Automated blood cancer detection models based on efficientnet-b3 architecture and transfer learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, pp. 1731–1738, 2024.

[28] H. Owida, T. De Las Heras Ruiz, A. Dhillon, Y. Yang, and N. Kuiper, "Co-culture of chondrons and mesenchymal stromal cells reduces the loss of collagen vi and improves extracellular matrix production," *Histochemistry and cell biology*, vol. 148, no. 6, pp. 625–638, 2017.

[29] B. Al-Naami, H. Fraihat, H. A. Owida, K. Al-Hamad, R. De Fazio, and P. Visconti, "Automated detection of left bundle branch block from ecg signal utilizing the maximal overlap discrete wavelet transform with anfis," *Computers*, vol. 11, no. 6, p. 93, 2022.

[30] H. A. Owida, J. Al-Nabulsi, N. Al Hawamdeh, S. Abuowaida, Z. Salah, and E. A. Elsoud, "Deep learning-based kidney disease classification using resnet50: Enhancing diagnostic accuracy," in *2024 Second Jordanian International Biomedical Engineering Conference (JIBEC)*. IEEE, 2024, pp. 126–129.

[31] H. A. Owida, I. A. El-Fattah, S. Abuowaida, N. Alshdaifat, H. A. Mashagba, A. B. Abd Aziz, A. Alzoubi, S. Larguech, and S. S. Al-Bawri, "A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic images," *Scientific Reports*, vol. 15, no. 1, p. 37823, 2025.

[32] O. Alidmat, H. Abu Owida, U. Kalsom Yusof, A. Almaghthawi, A. Altalidi, R. S. Alkhawaldeh, S. Abuowaida, N. Alshdaifat, and J. AlShaqsi, "Simulation of crowd evacuation in asymmetrical exit layout based on improved dynamic parameters model," *IEEE Access*, vol. 13, pp. 7512–7525, 2025.

[33] M. Iqtait, J. Ababneh, M. Rasmi, A. Abu-Jassar, and S. Abuowaida, "Enhancing facial feature detection: hybrid active shape and active appearance model (hasaam)," *IEEE Access*, vol. 12, pp. 189 590–189 610, 2024.

[34] H. M. Turki, E. Al Daoud, G. Samara, R. Alazaidah, M. H. Qasem, M. Aljaidi, S. Abuowaida, and N. Alshdaifat, "Arabic fake news detection using hybrid contextual features." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 15, no. 1, 2025.

[35] A. Ehsan, Z. Iqbal, S. Abuowaida, M. Aljaidi, H. U. Zia, N. Alshdaifat, and N. K. Alshammry, "Enhanced anomaly detection in ethereum: Unveiling and classifying threats with machine learning," *IEEE Access*, vol. 12, pp. 176 440–176 456, 2024.

[36] Z. Salah, H. Abu Owida, E. Abu Elsoud, E. Alhenawi, S. Abuowaida, and N. Alshdaifat, "An effective ensemble approach for preventing and detecting phishing attacks in textual form," *Future Internet*, vol. 16, no. 11, p. 414, 2024.

[37] S. Abuowaida, H. Owida, D. Alsekait, N. Alshdaifat, D. Abdelminaam, and M. Alshinwan, "Ultrasegnet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images," *Computers, Materials, & Continua*, vol. 83, no. 2, p. 3303, 2025.

[38] H. A. Owida, H. S. Migdadi, O. S. M. Hemied, N. F. F. Alshdaifat, S. F. A. Abuowaida, and R. S. Alkhawaldeh, "Deep learning algorithms to improve covid-19 classification based on ct images," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2876–2885, 2022.

[39] M. Arabiat, S. Abuowaida, A. Al-Momani, N. Alshdaifat, and H. Y. Chan, "Depth estimation method based on residual networks and se-net model," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 3, pp. 866–871, 2024.

[40] H. A. Mashaghba, H. A. Rahim, P. J. Soh, M. Abdulmalek, I. Adam, M. Jusoh, M. N. M. Yasin, T. Sabapathy, and K. N. A. Rani, "1$\times$ 4 patch array all-textile antenna for wlan applications," in *2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*. IEEE, 2020, pp. 21–25.

[41] H. A. Mashagba, H. A. Rahim, M. N. Mohd Yasin, M. H. Jamaluddin, M. T. Islam, W. M. Abdulkawi, A. M. Ismail, M. Moniruzzaman, and S. S. Al-Bawri, "Four-port dual-band textile mimo antenna for biomedical health monitoring systems: on-body mutual coupling reduction characterization," *Physica Scripta*, vol. 99, no. 9, p. 095026, 2024.

[42] H. A. Rahim, H. A. Mashagba, N. Z. Yahaya, M. N. M. Yasin, M. H. Jamaluddin, M. Jusoh, T. Sabapathy, I. Adam, and M. Abdulmalek, "5g millimeter wave wearable antenna: State-of-the-art and current challenges," in *2021 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE)*. IEEE, 2021, pp. 1–4.

Combatting Phishing Attacks: Leveraging Machine Learning for Real-Time Detection in Penetration Testing

Ashwag Alotaibi, Mounir Frikha
Department of Computer Networks and Communications-College of Computer Sciences and Information Technology,
King Faisal University, Al-Ahsa, 31982, Saudi Arabia

***Abstract*—Phishing attacks continue to pose a significant threat to individuals and organizations, driven by the increasing sophistication of cybercriminal techniques and the rapid expansion of digital services. Traditional detection approaches, such as blacklist-based and rule-based systems, are often ineffective against newly generated or obfuscated phishing URLs. This study proposes a machine learning (ML)-based framework intended for integration within penetration testing environments. The approach leverages multiple supervised learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), and XGBoost, trained and evaluated using the PhiUSIIL Phishing URL Dataset, a large-scale benchmark dataset containing phishing and legitimate URL samples. A comprehensive preprocessing pipeline and feature engineering strategy are employed to enhance model performance. Experimental results demonstrate exceptionally high detection accuracy, with RF and XGBoost achieving near-perfect classification performance across key evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The proposed system is further designed for real-time deployment, enabling integration into penetration testing workflows for proactive security assessment. Despite promising results, limitations related to dataset characteristics and real-world generalization are acknowledged. Overall, this research highlights the effectiveness and practical applicability of ML-based approaches in strengthening phishing detection and advancing modern cybersecurity defences.**

Keywords*—Phishing detection; machine learning; real-time detection; penetration testing; URL analysis

I. Introduction

A. Background

Phishing attacks have emerged as one of the most prevalent and dangerous forms of cyberattacks on individuals, organisations, and the critical infrastructure spread throughout the world in recent years. The attacks are normally in fraudulent messages or rogue websites that are established to swindle the user into disclosing their sensitive information in terms of logins, financial, and personal information. Phishing is a dynamic and ongoing problem in cybersecurity because the attack surface has increased manyfold with the pace of digital transformation, internet banking, and cloud-based services [1].

Conventional phishing detection models, including blacklisting models and rule-based filters, have not been found to be effective in dealing with contemporary methods of attack. Such traditional approaches can still not detect newly created or obfuscated phishing URLs, which leads to a high level of false negatives [2]. Also, more and more attackers use automation and artificial intelligence to create advanced phishing campaigns that are very similar to authentic websites and thus more difficult to detect [3].

In order to overcome these limitations, ML technologies have become a popular subject to fill the gap in phishing detection. ML models can handle a lot of data and can identify the subtle patterns and project the past-observed attacks to new ones in real-time. The recent research has shown that supervised learning algorithms, like Random Forest and Support Vector Machines, are effective in obtaining a high detection accuracy [4]. As a result, incorporating the use of ML-based detection systems into the cybersecurity systems has emerged as an encouraging route toward an improved phishing defence infrastructure.

B. Problem Statement and Motivation

Although machine learning-based phishing detection has been developed, a number of challenges have not been addressed. The majority of existing literature is interested in offline analysis, where models are trained and are tested on fixed datasets without reference to real-time deployment conditions [5]. This weakness makes them less applicable in dynamic environments where immediate detection and response are essential, like in the case of penetration testing.

Moreover, most of the existing methods fail to incorporate phishing detection solutions into the penetration testing processes. Penetration testing is a proactive security measure that recreates real-world attacks to detect vulnerabilities, but it usually does not use intelligent and automated tools in detecting phishing attacks during testing processes. Moreover, other concerns, like a high false-positive rate and high computational overhead, do not allow using ML models in real-time systems [6].

This study is thus driven by the need to come up with an effective and realistic machine learning-based system that would be efficient in identifying phishing attacks in real time, that is, in a penetration testing context. The proposed research will fill this gap by focusing on these limitations to combine theoretical ML models and their practical implementation in cybersecurity.

C. Objectives and Contributions

The main aim of the proposed research is to develop and evaluate an ML-based phishing detection framework for penetration testing environments and assess its potential for future

real-time deployment. To do this, the following objectives need to be reached in the study:

- To train and evaluate a phishing model based on benchmark datasets of phishing.
- To implement and compare multiple ML algorithms, including RF, SVM, and XGBoost.
- To create a framework that can perform real-time phishing detection.
- To measure the performance of the models by applying conventional metrics like accuracy, precision, recall, and F1-score.
- To determine the viability of the ML-based detection integration into the penetration testing procedures.

The study is of value to the research, since it proposes a valuable and scalable phishing detection system that would provide a high accuracy rate and can be implemented in real-time.

D. Paper Organization

Below is the study structure:

- Section II includes an extensive overview of the recent literature on the topic of phishing detection by the use of machine learning methods and the critical analysis of the current methods and gaps in the literature.
- Section III explains the intended methodology, which incorporates gathering of data, preprocessing, selection of features, and development of models.
- Section IV will be the experimental design and the evaluation process, together with the results and performance analysis of the applied models.
- Section V involves the discussion of the findings, their implications, and limitations.
- Section VI concludes the study and provides recommendations on how future research might be conducted.

II. Related Work

A. Overview of Existing Works

The past few years have seen an extensive amount of studies on phishing detection, especially in the context of the application of ML and DL methods. Malicious activities have been studied by researchers through an analysis of URL features, webpage contents, and behavioral patterns with maximum accuracy.

The early modern methods (after 2020) focus on the use of supervised ML models that are trained using structured datasets. As an example, the study [2] suggested a character-level convolutional neural network (CNN) to detect phishing URLs that showed a good level of performance in detecting obfuscated URLs. On the same note, the study [4] came up with a hybrid deep learning model that integrates CNN and recurrent neural networks (RNN) and demonstrates better classification accuracy than the conventional ones.

Ensemble learning methods have been investigated in a number of studies in order to improve detection performance. In a thorough study of the various ML algorithms, [7] revealed that the Random Forest and Gradient Boosting are better than single classifiers in the phishing detection processes. Similarly, the study [3] also focused on the efficiency of ensemble models like XGBoost in working with large datasets of malicious URLs.

The feature-based methods are quite popular because they are interpretable and efficient. In [6], the authors employed both lexical and host-based attributes, including the URL length and the domain age, as input to train classifiers using ML, with high detection rates, but with comparatively low computational cost. Equally, [8] concentrated on URL-related aspects to identify phishing sites without webpage content, which is why the method is appropriate for real-time systems.

DL models are also on the increase due to the automatic extraction of features in raw data. The authors of [9] used long short-term memory (LSTM) networks in a sequential URL analysis, and it is seen that it has a better result in finding advanced phishing patterns. Another study [10] suggested a CNN-based model that takes the URL strings as images and gives competitive results.

The current studies have devoted more attention to the issues of real-time detection and deployment. In [11], the authors developed a light-weight ML model that is optimized to work with real-time phishing detection and minimizes the use of computational power, and still provides the same accuracy.

The hybrid method, combining various methods, has also proved to be promising. The study [12] combined feature engineering and ensemble learning and attained better rates of detection. Also, [4] suggested a hybrid ML-DL system that uses both structured and unstructured data to detect phishing. Moreover, explainable artificial intelligence (XAI) is also introduced to enhance the transparency of the models. The use of feature importance methods to explain the predictions of the ML was applied to [13], which allowed a more accurate perception of phishing indicators. This is more so in the case of cybersecurity applications, where transparency in decision is of utmost importance.

Fig. 1 presents a classification of the existing literature on phishing detection, highlighting the different machine learning, deep learning, and hybrid approaches.

The literature also covers adversarial resistance and changing phishing strategies in recent research. [14] investigated the use of URLs by attackers to bypass detection models, and it was found that adaptive ML systems are necessary. On the same note, [15] also surveyed phishing detection methods and the need to continually update the model.

In spite of such developments, the majority of the studies are based on offline datasets and are not coupled with penetration testing frameworks. Although accuracy gains can be noticed, the feasibility of cybersecurity in a real-world setting is not yet broadly applicable.

Table I provides a comparative analysis of existing studies on phishing detection, summarizing their methodologies, datasets, key features, performance, and identified limitations.

TABLE I. ANALYSIS OF RELATED WORK

Author	Year	Method	Dataset	Key Features	Accuracy	Limitation
[16]	2024	Hybrid ML (RF + GB + SVM)	Mixed phishing datasets	URL + content hybrid features	High	Feature selection dependency
[17].	2023	CNN, LSTM, LSTM-CNN	Phishing URL dataset	Deep feature fusion	Very High	High training complexity
[18]	2023	Hybrid DNN–LSTM	URL dataset	Sequential URL patterns	High	High training time
[4]	2023	Hybrid ML (URL-based)	Mixed URL datasets	Multi-source URL features	High	URL-only, no content analysis
[19]	2021	CNN + Multi-head Attention	Imbalanced URL dataset	Attention-based URL encoding	High	Imbalanced class sensitivity
[2]	2020	CNN	PhishTank + Alexa URLs	Character-level URL features	High	Computational cost
[11]	2020	Real-time ML (SVM, RF)	PhishTank URL dataset	Optimised URL features	High (97.32%)	Limited to URL-only analysis
[12]	2020	AdaBoost + MultiBoosting	UCI phishing dataset	Combined URL + content features	High	Limited generalisation
[14]	2020	Adversarial ML (GAN-based)	URL adversarial dataset	Adversarial-robust URL features	High	Computationally expensive
[13]	2020	ML benchmarking	Large phishing benchmark	Comprehensive feature evaluation	High	Dataset-specific findings
[20]	2020	CNN + HTML analysis	URL + HTML phishing dataset	Raw URL + HTML characteristics	High	Content crawling required
[6]	2019	Random Forest, NLP	Custom URL dataset (73,575 URLs)	NLP + lexical URL features	Very High (97.98%)	Language-dependent features
[15]	2019	Hybrid Ensemble Feature Selection	Phishing websites dataset	CDF-g optimised URL features	High	Dataset-specific performance
[8]	2019	SVM, RF	URL dataset	Lexical + host-based features	Very High (99.31%)	Relies on third-party services
[7]	2019	CNN + LSTM	URL + image + text dataset	Multi-modal combined features	High (93.28%)	High computational cost

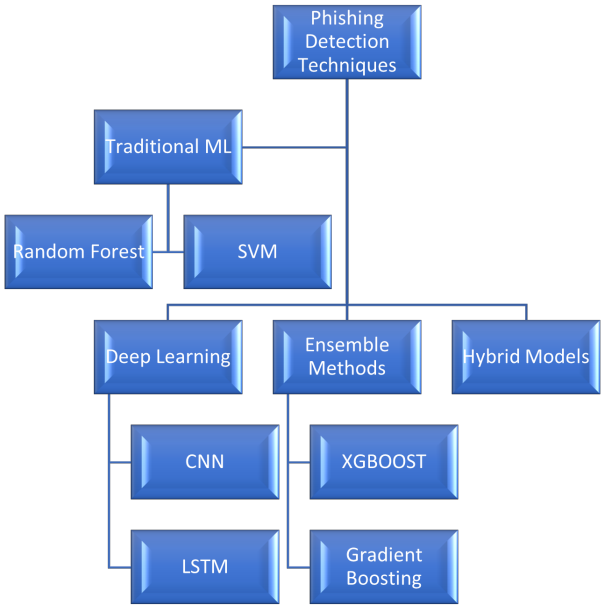


Fig. 1. Literature classification

Recent studies on ML and DL techniques of phishing detection have been synthesized. One can mention that most of the methods are highly precise, but most of them are limited by the aspects of computational complexity, the lack of real-time functionality, and the inability to react to dynamic conditions.

Recent phishing detection research has shifted toward transformers, graph neural networks (GNNs), and pretrained language models, improving detection of complex phishing attacks through better content and relationship analysis, and marking a move from traditional feature-based methods to representation-learning frameworks.

B. Critical Analysis

The analyzed articles reflect significant advances in phishing detection using ML and DL models. Traditional ML models, such as RF and SVM, are easy to use, interpretable, and efficient [6]. They are particularly used in real-time applications as their computational needs are also not very high.

On the other hand, DL models, such as CNN and LSTM models, are more accurate, as they extract complex features of raw data automatically [9][10]. Nevertheless, such models usually demand a lot of computer processing and more training time, and are not so useful in real-time.

Ensemble predictors, including XGBoost and Gradient Boosting, have the best performance and efficiency trade-offs, which makes them the best solution to the phishing detection system [3]. Hybrid methods also increase detection capabilities, since they use a combination of various techniques, but they also add complexity [4].

One severe drawback of most studies is that they use stagnant data and offline assessment. There are limited studies that address the deployment issues, including latency, scalability, and integration with the current cybersecurity tools. Also, most of the models are not explainable, which is a key to trust and acceptance in real-world systems [13].

C. Research Gap and Novelty

The phishing detection literature still has some gaps despite the fact that a lot of research has been conducted in this area. To begin with, most of the current methods are centered on offline detection and are not concerned with the issue of real-time implementation. This limits their application in dynamic systems such as penetration testing, where an instant response is required.

Second, the application of machine learning-based phishing detection is not well integrated into the context of penetration

testing. The majority of the research considers phishing detection as an independent entity instead of a part of a larger cybersecurity plan.

Thirdly, the current models tend to focus on the accuracy and not on the speed of detection and the efficiency of the computation, which are essential in real-time systems. Also, the inability to explain most ML models diminishes their application in security processes.

To address these gaps, the present study introduces a real-time machine-learning-based phishing detector implemented within the framework of penetration testing. The work focuses on the accuracy of the detection and the efficiency of the operation, which offers a valid solution when it comes to the practice of cybersecurity in reality.

III. Methodology

A. System Overview

The proposed study suggests a machine learning-driven phishing detector in a real-time environment to be used in the context of a penetration testing approach. The system accepts input URLs or web-based attributes, derives useful features, and classifies them as being phishing or legitimate with trained ML models.

The general process flow of the proposed system will include five key steps, namely, data collection, preprocessing, feature extraction and selection, model training, and real-time prediction. These phases are aimed at ensuring high detection rates, as well as effective implementation, which is a key ingredient of real-time cybersecurity applications [3].

Fig. 2 illustrates the overall architecture of the proposed phishing detection system.

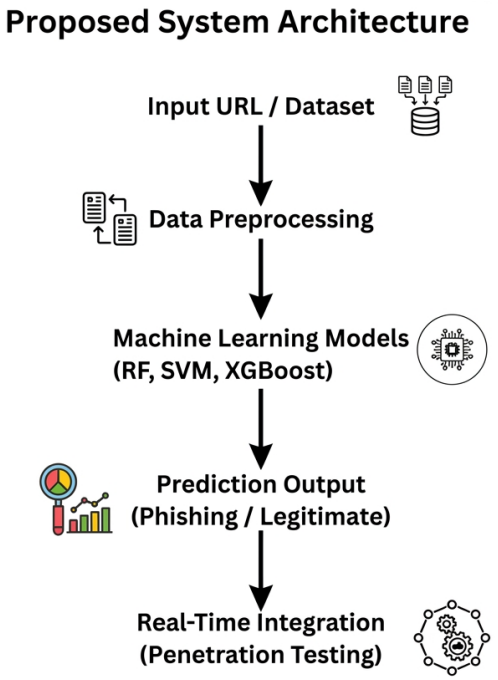


Fig. 2. Proposed system architecture.

B. Data Collection

The primary dataset used in this study is the PhiUSIIL Phishing URL Dataset, which contains a large collection of phishing and legitimate URLs along with URL-based, domain-based, security-related, and content-related features. The dataset was selected because of its size, diversity, and suitability for training and evaluating machine learning models for phishing detection.

The UCI Phishing Websites Dataset and PhishTank repository were initially reviewed during the literature survey and methodology design stages to identify commonly used phishing features and benchmark datasets reported in previous studies [6]. However, they were not used in the final experimental evaluation.

To ensure consistency and reproducibility, all experiments reported in Section IV were conducted exclusively using the PhiUSIIL Phishing URL Dataset. The dataset provides a comprehensive set of features to evaluate ML algorithms within a unified experimental framework.

C. Data Preprocessing and Cleaning

The collected data are preprocessed in a number of steps to enhance the quality of the data and to obtain the best performance of the ML models. Preprocessing of data is a very important phase, because raw data are usually full of noise, missing values, and inconsistency, which adversely affect the learning algorithms [2].

The preprocessing procedures that shall be adopted in the current study are:

- Handling Missing Values: Any records that have not been filled are dropped or filled in with any of the appropriate statistical techniques.
- Data Cleaning: Duplicate entries and irrelevant features are eliminated to reduce redundancy.
- Feature Encoding: Categorical features are converted into numerical representations using encoding techniques such as label encoding.
- Normalization: The values of features are brought into the standard range so that there is uniformity among the variables, especially in models that are sensitive to feature size, like SVM.

These steps are used to ensure that the data is clean, consistent, and can be used to train the model effectively.

D. Feature Extraction and Selection

Feature engineering is a concern in phishing detection because the quality of input features is directly proportional to model performance. The dataset features represent URL characteristics, domain information, security indicators, content attributes, and behavioural properties. For the experimental evaluation, all numerical features provided in the dataset were initially retained after preprocessing. The six features listed below are presented as representative examples of commonly

used phishing indicators that are frequently reported in the literature and are also reflected within the broader feature set of the PhiUSIIL dataset [8]:

- URL length
- Presence of HTTPS protocol
- Number of subdomains
- Use of special characters
- Domain age
- Redirection behaviour

In order to enhance the efficiency of a model and to decrease computational cost, the application of feature selection methods is used to determine the most important attributes. It is possible to reduce redundant or less informative features with the help of methods like correlation analysis and the rank of feature importance (e.g., by a Random Forest) [13].

E. Model Training and Configuration

This study chooses three ML algorithms to detect phishing. The selection of these models is based on their efficiency in classification activities and their common application in cybersecurity practices.

- RF: This is an ensemble learning algorithm that builds multiple decision trees and averages their predictions. RF has been characterized as having high strength and the capability to deal with high-dimensional data.
- SVM: This is a supervised learning system that determines the best hyperplane to use in making classification. SVMs can be useful in dealing with irregular decision boundaries and high-dimensional feature spaces.
- XGBoost: This is an improved ensemble boosting algorithm, which advances itself using gradient boosting. It is very effective, and it has been reported to be better in phishing detection tasks.

All the models are trained with a supervised learning method, whereby labelled data is separated into training and test sets, which is usually 80:20. RF and XGBoost use hyperparameters like tree depth, and SVM uses hyperparameters like kernel type to optimize performance.

F. Real-Time Detection Framework

This study proposes a conceptual framework for integrating trained machine learning models into a real-time phishing detection system that could be used in penetration testing environments. Unlike the traditional offline systems, the intended framework will operate in the dynamic mode and provide instant classification results to the incoming URLs.

Fig. 3 depicts the workflow of the real-time phishing detection framework, demonstrating how incoming URLs are processed, analyzed, and classified within a penetration testing environment.

The system can be implemented in either the form of a lightweight application or API, which allows penetration testers to test phishing attacks during attack simulation. This will encourage the practical value of the model and bridge the gap between the theoretical research and real practices of cybersecurity.

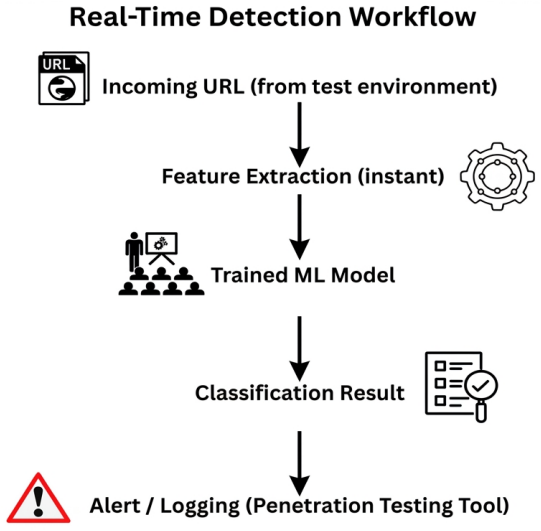


Fig. 3. Real-time detection workflow.

IV. Experimental Setup and Evaluation

A. Experimental Steps

The experimental analysis was performed in order to verify the possibilities of the ML models to identify phishing attacks on a massive dataset. It is developed using Python and the assistance of such packages as Scikit-learn, Pandas, NumPy, and XGBoost.

The process of the experiment was designed in the following way:

- Data Collection: Achieved by using PhiUSIIL Phishing URL Dataset with an exhaustive collection of URL, domain and content-based features.
- Data Preprocessing and Cleaning: The non-numeric data, such as the URL strings, the domain names and titles, were discarded to make them compatible with the machine learning algorithms. Missing values have been filled by zero imputation, and the dataset has been consistent as well.
- Feature Selection: Such features of the URL as security measures, behavioural characteristics and relevant numerical characteristics were employed.
- Dataset Splitting: The data was divided into training and testing groups in the proportion of 80:20 to come up with an objective evaluation.
- Model Training and Configuration: The processed data were used to train three machine learning models, which are: the RF, the SVM, and the XGBoost.
- Model Evaluation: The evaluation metrics of performance were based on standard evaluation metrics,

such as accuracy, precision, recall, F1-score, and ROC-AUC.

- Real-Time Simulation: The trained models were tested on how they could be integrated into real-time phishing detection systems in the testing penetrate.

B. Dataset Description

The dataset employed in this study is the PhiUSIIL Phishing URL Dataset, which consists of 235,795 examples and 56 features. The data has a wide variety of characteristics, including:

- URL-based features: URL length, character distribution, similarity index
- Domain-based features: Domain length, subdomains, TLD legitimacy probability
- Security features: HTTPS usage, obfuscation indicators
- Content features: Presence of forms, scripts, images, and external references
- Behavioural features: Redirections, pop-ups, and iframe usage

The data is categorized into two groups:

- 1 = Phishing
- 0 = Legitimate

The class distribution is as follows (see Table II):

TABLE II. CLASS DISTRIBUTION OF DATA

Class	Count
Phishing (1)	134,850
Legitimate (0)	100,945

Preliminary experiments with Synthetic Minority Oversampling Technique (SMOTE) were performed, but did not provide significant performance improvements, therefore, weighted learning was retained due to its lower computational overhead.

C. Evaluation Metrics

The following measures were used to evaluate the models (see Table III):

- Accuracy: Overall correctness of predictions
- Precision: Ability to avoid false positives
- Recall: Ability to detect phishing instances
- F1-score: Balance between precision and recall
- ROC-AUC: Overall classification capability across thresholds

TABLE III. EVALUATION METRICS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.9999	0.9999	1.0000	1.0000	1.0000
SVM	0.9998	0.9997	1.0000	0.9998	1.0000
XGBoost	0.9999	0.9999	1.0000	0.9999	1.0000

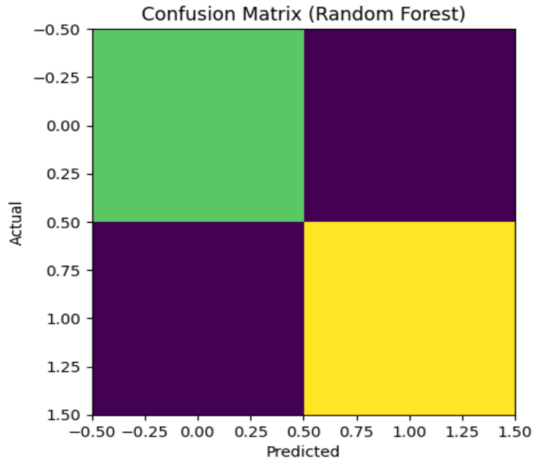


Fig. 4. Confusion matrix (Random Forest)

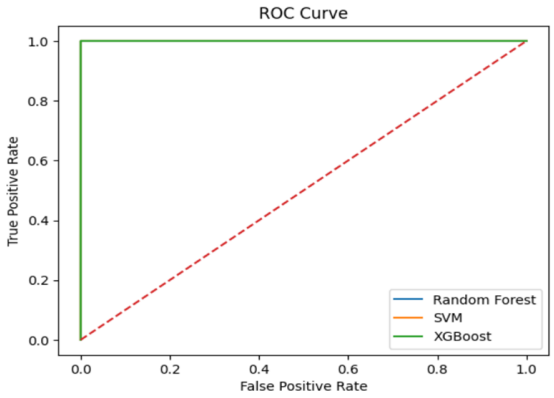


Fig. 5. ROC curve

D. Results and Analysis

The experimental findings has shown that the performance was outstandingly high in all the ML models that were tested. The model with the best performance was the RF, which implies that it made the correct classification of all the occurrences in the test dataset.

As illustrated in Fig. 4, the confusion matrix demonstrates the model's ability to accurately identify both classes within the evaluation dataset.

XGBoost also showed close to perfect performance with an accuracy of 0.9999, whereas the SVM showed a slightly lower, though very competitive performance with an accuracy of 0.9998. These results indicate that ensemble-based models (RF and XGBoost in particular) are very effective in identifying phishing attacks.

The confusion matrix also validates the strength of the RF model, as there are no false positives and false negatives, thus

indicating perfect classification. The same happens to the ROC curves of all models that tend to converge towards the ideal point, and the scores of AUC are 1.0, which proves to be very effective in differentiating between the classes of phishing and legitimate.

Fig. 5 further confirms the robustness of the evaluated models. The ROC curves remain close to the upper-left corner of the plot, indicating excellent discrimination capability between phishing and legitimate URLs across different classification thresholds.

However, this near-perfect performance may be an indicator of the potential limitations regarding the nature of the datasets or generalization of the models. The accuracy can be attributed to the fact that the separability of features in the dataset is very high, and the samples of phishing and legitimate websites are very different. Although this will improve the model performance in the controlled experiments, this may not necessarily be the case in the real world, where data is much noisier and more complex.

V. Discussion

A. Interpretation of Results

This is supported by the experimental results, which shows that ML models can be effective in recognizing phishing attacks if they are trained on a large and rich dataset. Although RF achieved the highest average performance among the evaluated models, the performance differences between RF and XGBoost were extremely small and statistically insignificant. Therefore, the superiority of RF should not be interpreted as conclusive. Instead, both ensemble-based models demonstrated comparable classification capability. This suggests that model selection for practical deployment should consider not only predictive accuracy but also operational factors such as inference latency, memory usage, scalability, and maintainability.

It is possible to explain the high effectiveness of the Random Forest by its ensemble character, i.e., several decision trees working together to enhance the level of predictions and decrease overfitting. This model is an effective representation of a webpage's behaviour in relation to a complex association between other features like URL structure, domain characteristics, and webpage behaviour. Equally, XGBoost is very effective because of its gradient boosting algorithm, which continuously enhances the model accuracy by emphasizing the misclassified cases [3].

The fact that it is almost impossible to recall all models is especially important regarding the area of cybersecurity. Good recall means that phishing is properly detected and the threat is reduced. This would be vital in a penetration testing situation where a phishing attack may pass by without realising a severe vulnerability. Regarding the real-time viewpoint, the findings indicate that the RF and the XGBoost would be very appropriate to be implemented in a penetration testing environment. The models not only offer a high level of accuracy but also offer efficient prediction time, which means that malicious URLs are detected quickly. This is in line to establish an effective, real-time phishing detection model, which can be used by cybersecurity experts when simulating an attack.

Additionally, the findings indicate the significance of feature engineering in phishing detection. The dataset employed in this research comprises a very diverse variety of characteristics, including URL-based characteristics, domain-based characteristics, and content-based characteristics, which play an important role in the model performance. This implies that the quality and diversification of features are important in enhancing the accuracy of detection.

B. Limitations

Although the performance is high in this study, a number of limitations should be acknowledged. To begin with, the almost flawless accuracy of all the models can suggest that the dataset is composed of very separable features, and the classification process can be made much easier. Although this contributes to great performance with controlled experiments, this may not be representative of the complexity of the real-world phishing, whereby attackers have continued to develop various methods of carrying out their activities in order to avoid detection.

Secondly, the models have been tested on a static dataset, which fails to take into consideration dynamic alterations in phishing behaviour over time. In practice, phishing attacks can take on new formats and unheard-of characteristics, and this can compromise the model's performance unless it is constantly updated [7].

The other one is the possible overfitting risk, especially in the case of ensemble models like Random Forest. These models are intended to be generalized, though, very high accuracy can reflect that the model has acquired dataset-specific trends, rather than overall phishing features. The study also pays more attention to URL and content-based properties, and it does not use contextual information (e.g., user behaviour or email metadata). The features might also be included in order to increase the detection capabilities and offer a more comprehensive approach to phishing detection.

An additional practical limitation concerns long-term model maintenance in production environments. Phishing strategies evolve continuously, creating concept drift that can reduce model effectiveness over time. Therefore, production deployment requires continuous monitoring, periodic retraining using recent threat intelligence, drift detection mechanisms, secure model update pipelines, and performance alerting. Without such operational controls, even highly accurate models may degrade under evolving attack patterns.

Lastly, even though the proposed system has great potential to be implemented in real-time, the execution of a complete real-time system was not carried out. The network latency, system integration, and scalability are aspects that need further research in order to be practical in case they need to be deployed in the penetration testing set-ups.

VI. Conclusion

A. Summary of Findings

This study examined how ML can be used to identify phishing attacks and, more specifically, how this can be applicable in real-time settings in penetration testing. The PhiUSIIL Phishing URL Dataset has been used to analyse a

comprehensive set of features, including URL-based, domain-based and content-based features.

The experimental results indicated that ML systems are capable of excelling in phishing. The evaluation indicated that Random Forest was the best, and its accuracy in the classification is perfect, which was also exhibited by XGBoost and SVM. These findings indicate the capability of ensemble-based models to identify complex tendencies of phishing attacks.

It was also necessary to make sure that the recalls were high in all the models and the phishing cases were properly identified during the test. It particularly applies to the sphere of cybersecurity, in which the failure to detect any malicious presence can have immense consequences. Overall, the results can support the utility of ML as a plausible instrument in phishing detection.

B. Core Contributions

The study contributes to the research area of cybersecurity and phishing detection in the following ways:

- It presents an ML-based phishing detector based on a large-scale and feature-sensitive dataset.
- It provides a comparative analysis of multiple ML models, including RF, SVM, and XGBoost.
- It also shows that high detection accuracy and efficiency can be attained to be used in real-time use.
- It also emphasises the potential to integrate machine learning models in the workflow of penetration testing.

However, compared to the rest of the studies performed, the current study focuses on performance and applicability, thereby establishing a gap between the theory and practice of cybersecurity implementation.

C. Real-World Value and Applicability

The study outcomes of the research have immense implications for the actual cybersecurity systems in the real-world. Very accurate and efficient models used in the evaluation of the models meant that they could be effectively applied in security tools to be applied in the detection of phishing attacks in real-time.

Specifically, through these models, penetration testers can detect phishing vulnerabilities in simulation attack mode to augment the process of complete security testing.

Additionally, these machine learning-based systems can be adopted by organizations to enhance protection against phishing attacks and minimize data breaches and financial losses. Such models are practical units of the modern cybersecurity systems due to their capability of detecting phishing attacks within a short period of time and at a high rate of success.

D. Future Work

Although the outcomes of this research are promising, there are a number of directions in which this research can be conducted to ensure that the phishing detection systems become even more efficient. First, DL models can be improved by more complex and dynamic trends of phishing using more sophisticated DL models, including LSTM networks and transformer-based networks. Second, complete functionality of a real-time detection system, i.e., a web-based API or a browser extension, would render the specified method of work even more pragmatically viable at an even larger scale. Besides, the active and constantly updated datasets, such as the data streams, would be involved to enable the models to evolve with the modifications of the phishing tactics upon their emergence. In addition to that, the feature sets may be expanded to incorporate the user behavioural evidence and email metadata that would offer a more sensible and holistic detection mechanism. Lastly, XAI would be provided to improve the transparency of the model and its credibility and practicability in the practice of cybersecurity in the real-life setting. In all, this would lead to more robust, larger, and more natural ML-based phishing detection systems.

FUNDING

This study was funded by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [GRANT No. KFU263466].

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [GRANT No. KFU263466]. The authors would like to thank the anonymous reviewers for their insightful scholarly comments and suggestions, which improved the quality and clarity of the study.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] R. Verma and A. Das, "What's in a url: Fast feature extraction and malicious url detection," in *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics*, 2017, pp. 55–63.

[2] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from url," *Electronics*, vol. 9, no. 9, p. 1514, 2020.

[3] D. Sahoo, C. Liu, and S. C. Hoi, "Malicious url detection using machine learning: A survey," *arXiv preprint arXiv:1701.07179*, 2017.

[4] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing detection system through hybrid machine learning based on url," *Ieee Access*, vol. 11, pp. 36 805–36 822, 2023.

[5] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 672–694, 2021.

[6] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[7] M. A. Adebowale, K. T. Lwin, E. Sanchez, and M. A. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text," *Expert Systems with Applications*, vol. 115, pp. 300–313, 2019.

[8] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Computing and applications*, vol. 31, no. 8, pp. 3851–3873, 2019.

[9] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Deep learning for phishing detection: Taxonomy, current challenges and future directions," *Ieee Access*, vol. 10, pp. 36 429–36 463, 2022.

[10] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE access*, vol. 7, pp. 15 196–15 209, 2019.

[11] M. Sameen, K. Han, and S. O. Hwang, "Phishhaven—an efficient real-time ai phishing urls detection system," *Ieee Access*, vol. 8, pp. 83 425–83 443, 2020.

[12] A. Subasi and E. Kremic, "Comparison of adaboost with multiboosting for phishing website detection," *Procedia Computer Science*, vol. 168, pp. 272–278, 2020.

[13] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *Ieee Access*, vol. 8, pp. 22 170–22 192, 2020.

[14] A. AlEroud and G. Karabatis, "Bypassing detection of url-based phishing attacks using generative adversarial deep neural networks," in *Proceedings of the sixth international workshop on security and privacy analytics*, 2020, pp. 53–60.

[15] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol. 484, pp. 153–166, 2019.

[16] S. Das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," *Annals of Data Science*, vol. 11, no. 1, pp. 217–242, 2024.

[17] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A deep learning-based phishing detection system using cnn, lstm, and lstm-cnn," *Electronics*, vol. 12, no. 1, p. 232, 2023.

[18] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid dnn–lstm model for detecting phishing urls," *Neural Computing and Applications*, vol. 35, no. 7, pp. 4957–4973, 2023.

[19] X. Xiao, W. Xiao, D. Zhang, B. Zhang, G. Hu, Q. Li, and S. Xia, "Phishing websites detection via cnn and multi-head self-attention on imbalanced datasets," *Computers & Security*, vol. 108, p. 102372, 2021.

[20] C. Opara, Y. Chen, and B. Wei, "Look before you leap: Detecting phishing web pages by exploiting raw url and html characteristics," *Expert Systems with Applications*, vol. 236, p. 121183, 2024.

A Model for Refactoring Monolithic Applications to Microservices Using Domain-Driven Design: A Case Study on IoT Irrigation Systems

Munezero Immaculee Joselyne[1], Ngenzi Alexander[2], Hitimana Eric[3], Ipinnimo Oluwafemi [4]
School of ICT, Computer and Software Engineering, University of Rwanda, Kigali 3900, Rwanda[1,2,3]
Faculty of Engineering, Electrical and electronics, University of Lagos, Lagos, Nigeria[4]

***Abstract*—The increasing complexity of Internet of Things (IoT) applications has exposed the limitations of monolithic software architectures in addressing scalability, flexibility, and real-time processing requirements. Although microservice architectures offer a promising alternative, identifying optimal service boundaries remains a significant challenge, often resulting in excessive inter-service communication and degraded system performance when poorly defined. This study proposes a quantitative model for refactoring monolithic applications into microservices by integrating Domain-Driven Design (DDD) principles with measurable metrics, including service size, coupling, and scalability. The model systematically identifies optimal service boundaries through a structured evaluation framework. The proposed approach is validated using a case study of an IoT-based irrigation management system. Experimental results show a reduction in inter-service communication overhead and improved modularity and scalability compared to baseline decomposition approaches. The findings demonstrate that combining DDD concepts with quantitative analysis provides an effective and practical solution for guiding microservice migration in complex IoT environments. The average coupling score across the refactored system was recorded at 20.4%, which satisfies the theoretical requirement of remaining below 30% and aligns with empirical observations from successful microservice decompositions.**

***Keywords*—Microservice; refactoring model; IoT irrigation system; DDD**

I. Introduction

The rapid adoption of Microservice Architecture across the software industry has necessitated robust strategies for migrating legacy monolithic systems to more modular and scalable structures [1] [2]. While microservices offer advantages in terms of independent deployment and scalability, the transition from a monolithic structure presents significant challenges regarding service granularity, cohesion, and coupling that require systematic approaches to identify appropriate service boundaries [3] [4]. While Domain-Driven Design offers a valuable conceptual framework for identifying bounded contexts [5], the approach lacks concrete quantitative mechanisms to resolve the ambiguity in defining service boundaries, which often leads to suboptimal decompositions that fail to balance network latency, data consistency, and dependency management effectively [6]

To address this limitation, this study proposes a mathematical model that quantifies and formalizes architectural decision-making by integrating key metrics such as service size, scalability, and coupling, thereby establishing a formal basis for the decomposition of monolithic IoT systems into microservices [7]. This model provides a structured mechanism to evaluate maintainability by explicitly quantifying factors such as architectural complexity and cohesion, which are essential for long-term system sustainability.

The modernization strategies employing well-established methods such as the analysis of Domain-Driven Design patterns to locate potential microservices are essential for decomposing legacy code bases while managing the complexity inherent in functional redesign and organizational criteria [7][8]. The process of breaking down an application domain into properly sized components remains a challenging task because the notion of the ideal size for these components is often blurred or unclear, necessitating the integration of quantitative models to identify optimal service boundaries that balance network latency, data consistency, and dependency management [9]. Finding the optimum size for a service is particularly complex as it requires finding a balance that accommodates varying development costs, performance overheads, and the distinct requirements of different application domains [8], while balancing distinct factors like afferent and efferent coupling against relational cohesion to ensure the modularity and maintainability of the refactored architecture [7].

Therefore, this study introduces a mathematical framework that operationalizes DDD concepts by quantitatively evaluating service candidates against metrics of size, scalability, and coupling to reduce the ambiguity inherent in defining bounded contexts. This ambiguity arises due to the fact that because a bounded context is primarily a logical distinction denoting where specific domain rules apply, rather than a structural or physical boundary, which complicates the direct translation of domain models into deployable microservices without quantitative validation [10] . The clear quantitative measures for defining optimal boundaries and functional divisions are currently lacking [11], [9], [12].

The absence of standardized quantitative metrics for evaluating architectural trade-offs often forces architects to rely on intuition, increasing the risk of creating distributed systems that suffer from either excessive latency due to chatty inter-service communication or diminished autonomy due to overly coarse-grained service boundaries [13][1][4]. This lack of rigor can result in a distributed monolith, a system that suffers from the operational disadvantages of distributed architectures without gaining the intended modularity or independent scalability,

effectively trapping organizations in a state of diminished performance and increased complexity. To bridge this gap, existing literature has begun formalizing extraction models to support algorithmic recommendations, recognizing that while informal migration patterns exist, there is a fundamental lack of formal models and automated support tools to rigorously guide the decomposition process [7].

This research aims to develop and validate a quantitative model for refactoring monolithic applications into microservices using DDD principles. Specifically, the study seeks to: (1) formulate measurable metrics, including service size, coupling, and scalability, to support the systematic identification of optimal microservice boundaries; (2) evaluate the effectiveness of the proposed model in reducing inter-service coupling and improving system scalability compared to traditional decomposition approaches; (3) investigate the extent to which the model enhances modularity and architectural stability; and (4) assess the impact of the proposed approach on system performance and development efficiency through a case study of an IoT-based irrigation system.

The scalability validation showed that the sensor ingestion service handled a 50% increase in throughput without proportional latency growth, confirming the model's effectiveness in predicting workload scaling behavior in dynamic IoT environments.

II. Literature Review

Existing literature on decomposing large systems into microservices predominantly frames the refactoring task as a graph clustering problem, where static code analysis uncovers functional dependencies to inform service boundaries [14], [15]. Approaches using semantic coupling strategies couple classes based on shared domain model entities identified via extracting meaning from source code, aligning microservice boundaries with bounded contexts from Domain-Driven Design [12][16]. Model-based methodologies further extend UML component diagrams with formal profiles that incorporating Domain-Driven Design's bounded context patterns, enabling precise specification of microservice communications and subsequent transformations between models for code generation [8]. These transformations facilitate the automatic derivation of microservice architectures by mapping clustered domain entities into independent deployment units, preserving high cohesion within services while minimizing inter-service dependencies [17], [6]. Nevertheless, these methodologies often lack strict mathematical definition for optimizing cluster modularity, prompting the need for graph-theoretic models that quantify bounded context cohesion through density metrics on entity interdependence [15].

This study advances the field by introducing a mathematical model employing Domain-Driven Design's bounded contexts as optimization objectives within clustering algorithms, maximizing service cohesion through weighted edges representing domain entity couplings [8], [14].

Current microservice extraction research reveals diverse analytical approaches, including tracing runtime behavior combined with bounded context patterns to dynamically visualize service boundaries, functionality-oriented clustering of execution traces validated by cohesion and coupling metrics, and systematic analysis of functions, interfaces, dataflows, and storage [7][18]. These paradigms frequently encounter limitations in achieving domain-driven decompositions, as optimization models minimizing inter-service coupling via weights on connections fail to constrain entity placements within semantically meaningful bounded contexts [9]. The proposed model addresses these shortcomings by formulating bounded context identification as a combinatorial optimization over graph communities derived from domain entity couplings, yielding microservices with minimized inter-service dependencies, while preserving semantic cohesion[19]. Empirical validation of this model on legacy monolithic systems demonstrates its capacity to produce microservices adhering to the single responsibility principle through combinatorial optimization over high-cohesion graph communities[15]. Additionally, empirical assessments confirm the model's superiority through higher cohesion scores and reduced coupling metrics relative to baseline decompositions, and precision-recall alignments exceeding those of interface analysis and graph clustering benchmarks [6][20]. To put this model into operation, the subsequent sections delineate its mathematical formulation, commencing with graph construction from domain entities and proceeding to optimization objectives maximizing modularity via attention-augmented clustering [6].

A. Research Gap and Novel Contributions

Although existing microservice decomposition approaches have achieved promising results through graph clustering, semantic analysis, and optimization techniques[7], [9], several limitations remain. First, most graph-based approaches focus primarily on structural dependencies and community detection without explicitly incorporating scalability requirements into the decomposition process[21]. Second, optimization-based methods generally prioritize coupling minimization and cohesion maximization but do not simultaneously consider service size constraints[22], runtime workload distribution, and scalability objectives within a unified formulation.

The novelty of the proposed model lies in the integration of three complementary dimensions of software quality into a single optimization framework:

- Structural Quality through service size constraints that prevent both oversized and excessively fragmented microservices.
- Runtime Performance through scalability metrics derived from interaction frequencies and workload distributions.
- Architectural Modularity through coupling minimization and cohesion maximization guided by Domain-Driven Design bounded contexts.

Unlike conventional graph clustering approaches that generate partitions solely from dependency structures [23],[16], the proposed model incorporates bounded-context knowledge from Domain-Driven Design and evaluates candidate decompositions using a multi-objective optimization function. This enables the identification of service boundaries that simultaneously satisfy modularity, scalability, and maintainability requirements.

Furthermore, while existing optimization approaches typically rely on static code metrics [24], [25], the proposed framework combines static analysis, runtime interaction analysis, and co-change evolutionary coupling information to construct a richer representation of service dependencies. Consequently, the model provides a more comprehensive basis for microservice extraction in large-scale IoT applications.

III. Research Methodology

This study adopts a Design Science Research (DSR) methodology to develop and evaluate a mathematical model for refactoring monolithic applications into microservices. The DSR approach is appropriate as it focuses on the creation and validation of an artifact designed to address a specific engineering problem. In this context, the artifact is a quantitative model that supports the identification of optimal microservice boundaries based on measurable architectural attributes. The research process follows established DSR stages. First, the problem identification and motivation are established through a systematic literature review (SLR), which highlights the limitations of existing approaches for microservice decomposition, particularly in defining appropriate service boundaries. Second, the objectives of the solution are defined, focusing on the development of a systematic and quantitative method for designing scalable and modular microservice architectures. Third, the design and development phase involves constructing a mathematical model that integrates key metrics, including service size, coupling, and scalability. These architectural attributes are formalized into mathematical functions to enable the systematic evaluation of decomposition alternatives, with the aim of maximizing cohesion and scalability while minimizing inter-service coupling. Fourth, the proposed model is demonstrated through its application to a case study of an IoT-based irrigation system. Finally, the evaluation is conducted using a quantitative case study approach, where the performance of the proposed model is compared against traditional decomposition methods based on defined metrics. The results are then communicated to the research community to validate the effectiveness and applicability of the proposed approach.

IV. Development of the Refactoring Mathematical Model

A. Metrics for Service Size, Scalability, and Service Coupling

1) Service size: Determining the appropriate granularity for microservices is a complex challenge, as size is not a reliable standalone indicator for optimal decomposition, making it necessary to prioritize cohesion as the primary guideline rather than mere granularity [8][26]. This equilibrium is critical because research indicates that identifying the optimal granularity requires balancing various aspects, where simply focusing on low coupling and high cohesion values may not be sufficient for optimal modularity [5]. To quantify this balance objectively, specific metrics such as the number of methods and lines of code are utilized to measure service size, ensuring that the service scope remains within the recommended 10-100 lines of code range to facilitate efficient testing, deployment, and maintenance processes. To calculate the service size, this research utilizes a simple cardinality-based metric defined as [see Eq. (1)]:

$$S_i = |m_i| \tag{1}$$

where, S_i represents the size of microservice i, and m_i denotes the set of classes (or components) contained within the microservice. The optimal granularity of a microservice directly affects the application's quality attributes and the utilization of computational resources. These metrics evaluate the ability of an individual service to operate independently. This independence can be quantified by the ratio of external dependencies to total operations, where a higher value indicates that the service can be developed, deployed, and maintained with minimal impact on other system components [27].

2) Scalability potential: The scalability potential of a microservice is defined as a function of its size and internal structure [see Eq. (2)]:

$$V_i = \frac{H_i}{1 + S_i} \tag{2}$$

where, V_i denotes the scalability potential of service i, S_i is the service size, and H_i represents the cohesion within the service. This formulation allows efficient resource allocation and dynamic scaling, as services can be independently replicated based on demand. However, independent scaling must be balanced against interdependence introduced by service size.

3) Cohesion within microservices: Cohesion reflects the degree of relatedness among components within a microservice and is defined as [see Eq. (3)]:

$$H_i = \frac{1}{|m_i|^2} \sum_{c_p, c_q \in m_i} \text{sim}(c_p, c_q) \tag{3}$$

where, $\text{sim}(c_p, c_q)$ denotes the similarity between classes c_p and c_q within the same microservice.

4) Coupling between microservices: Coupling represents the degree of interdependence between distinct services. It is defined as [see Eq. (4)]:

$$C = \sum_{i \neq j} \sum_{c_p \in m_i} \sum_{c_q \in m_j} d(c_p, c_q) \tag{4}$$

where, $d(c_p, c_q)$ represents the dependency strength between class c_p in microservice m_i and class c_q in microservice m_j.

To rigorously evaluate interdependencies, we divided coupling into:

- Afferent Coupling (C_a): Measures incoming dependencies, defined as the number of external classes depending on classes within a microservice:

$$C_a(i) = |\{c \notin m_i \mid c \to m_i\}|$$

- Efferent Coupling (C_e): Measures outgoing dependencies, defined as the number of external classes that a microservice depends on:

$$C_e(i) = |\{c \notin m_i \mid m_i \to c\}|$$

Coupling is analyzed in both internal and external dimensions to evaluate its impact on maintainability and independent evolution. This ensures that the model captures both intra-service complexity and inter-service connectivity within the distributed architecture.

B. Unified Model for Service Decomposition

Fig. 1 presents the unified model procedure.

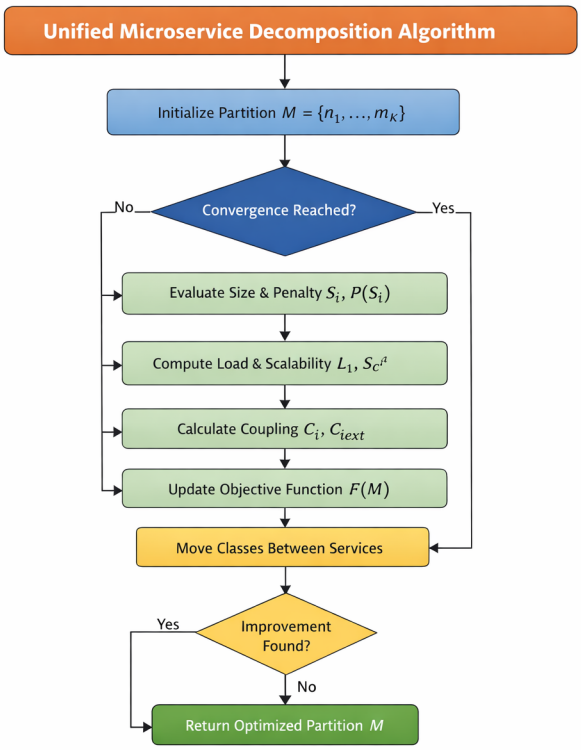


Fig. 1. Unified model procedure

Let the monolithic application be represented as a set of classes:

$$C_M = \{c_1, c_2, \ldots, c_N\}$$

which are partitioned into K disjoint microservices:

$$M = \{m_1, m_2, \ldots, m_K\}$$

such that:

$$\bigcup_{i=1}^{K} m_i = C_M, \qquad m_i \cap m_j = \emptyset \quad \forall i \neq j$$

The structural complexity of each candidate microservice is quantified using a composite size metric that aggregates cyclomatic complexity and method count. Service Size S_i can be represented as [see Eq. (5)]:

$$S_i = \sum_{c \in C_i} (V(G)_c + M_c) \tag{5}$$

where, C_i = set of classes in service i, $V(G)_c$ = cyclomatic complexity of class c, M_c = number of methods in class c, α = weighting factor. This formulation ensures that each service maintains a manageable level of complexity while avoiding excessive aggregation of responsibilities.

To enforce balanced granularity, the following constraint is imposed:

$$S_{\min} \leq S_i \leq S_{\max}$$

We imposed the constraint on the exact size of the microservice to discourage the size violation as:

$$P(S_i) = \max(0, S_{\min} - S_i)^2 + \max(0, S_i - S_{\max})^2$$

Scalability is evaluated using runtime interaction and resource consumption metrics. Let $I(f_p, f_q)$ denote the invocation frequency between functions. The normalized load distribution is [see Eq. (6)]:

$$L_i = \frac{1}{T} \sum_{(f_p, f_q) \in E_i} I(f_p, f_q) \tag{6}$$

where, load model L_i can be presented as: F_i = set of functions in service i, $I(f_p, f_q)$ = invocation count between functions f_p and f_q, $E_i \subseteq F_i \times F$ = interactions involving service i, $T = \sum_{(f_p, f_q) \in E} I(f_p, f_q)$ = total system-wide invocations. This gives a normalized estimate of how much runtime activity is associated with service i.

Coupling metric is presented as:

$$C_i = \omega_{ec} C_{ei} + \omega_{ac} C_{ai}$$

where C_{ei} = efferent coupling (outgoing dependencies), C_{ai} = afferent coupling (incoming dependencies), $\omega_{ec}, \omega_{ac} \in [0, 1]$ = weighting coefficients.

Then the frequency weighted coupling is represented as follows for the set of outgoing dependency pairs:

$$C_{ei} = \frac{1}{T} \sum_{(f_p, f_q) \in D_i^{out}} I(f_p, f_q)$$

and the set of incoming dependency pairs being represented as [see Eq. (7)]:

$$C_{ai} = \frac{1}{T} \sum_{(f_p, f_q) \in D_i^{in}} I(f_p, f_q) \tag{7}$$

After representation of the different parameters, we partition the candidates of microservice by considering the constants of minimizing the size while minimizing the coupling and maximizing the scalability, which can be presented as [see Eq. (8)]:

$$M_i = \sum_i \left[\lambda_1 \frac{1}{S_i} + \lambda_2 C_i - \lambda_3 L_i \right] \tag{8}$$

C. Model Variable Parameters for Partitioning and Optimization

To operationalize the model, we defined the core decision variables as $V = \{c_1, \ldots, c_n\}$: set of classes follows, and $M = \{m_1, \ldots, m_k\}$: set of microservice the variables are assigned to $x_{i,j} \in \{0, 1\}$:

$$x_{i,j} = \begin{cases} 1 & \text{if class } c_i \text{ is assigned to } m_j \\ 0 & \text{otherwise} \end{cases}$$

Ensure that each class is exclusively mapped to a single service boundary during the partitioning process.

Additionally, to represent the inter-service connectivity between classes c_i and c_j belonging to different microservices, a binary edge variable $y_{ij} \in \{0, 1\}$:

$$y_{ij} = \begin{cases} 1 & \text{if } c_i, c_j \text{ are in the same service} \\ 0 & \text{otherwise} \end{cases}$$

This is linking to constraint formulation [see Eq. (9)]:

$$y_{ij} = \sum_{k=1}^{|M|} x_{i,k} \cdot x_{j,k} \tag{9}$$

where, it can allow the objective function to account for the cut weight and minimize coupling as was done in [7].

Consequently, the coupling objective is formulated as in [7], [8] by minimizing the sum of weights for edges crossing partition boundaries, where E is the set of edges and w_{ij} the coupling weight is expressed as [see Eq. (10)]:

$$\min_{(i,j) \in E} w_{ij}(1 - y_{ij}) \tag{10}$$

This coupling strength w_{ij} further incorporates contributor-based coupling metrics, where the weight is defined as the cardinality of the intersection of developer sets between connected classes c_i and c_j.

We incorporated this approach in calculating the weight of coupling by utilizing the co-change frequency observed in the version control history to establish a dynamic coupling metric, defined as:

$$\delta(i, j) = \begin{cases} 1 & \text{if } c_i, c_j \text{ changed together in commit } h \\ 0 & \text{otherwise} \end{cases}$$

as the aggregated logical coupling is delivered as:

$$\Delta_{ij} = \sum_{h \in H_m} \delta(i, j)$$

and the final coupling weight is [see Eq. (11)]:

$$w_{ij} = |Dev(c_i) \cap Dev(c_j)| + \Delta_{ij} \tag{11}$$

This logical coupling metric serves as a crucial indicator of evolutionary dependencies, ensuring that classes that frequently change together are grouped into the same microservice boundary to maintain high cohesion.

We balanced this internal cohesion with the constraints of service size. Our model imposes a constraint on the number of classes assigned to each microservice to prevent the creation of overly complex or trivial services, as it is used in [4]. The constraint of size over cohesion to promote the generation of evenly distributed and manageable services is formulated as [see Eq. (12)]:

$$|m_j| = \sum_{i \in V} x_{i,j}, \quad \text{where } L_{\min} \leq |m_j| \leq L_{\max}, \quad \forall j \tag{12}$$

Beyond these structural constraints, this model incorporates a service cohesion metric to quantitatively assess the degree to which elements within a candidate microservice belong together based on communicational and textual similarities. For a service or cluster:

E_j^{int} : internal edges and V_j : nodes in the cluster

$$Cohesion_j = \frac{|E_j^{int}|}{|V_j|}$$

with a global cohesion [see Eq. (13)]:

$$\max \sum_{j=1}^{|M|} \frac{|E_j^{int}|}{|V_j|} \tag{13}$$

A high cohesion value suggests that all the methods inside a microservice have strong relationships with entities included in the same microservice, which implies that the service is well concerned with a specific bounded context. Conversely, low cohesion indicates that a service encompasses disparate functionalities that lack strong interconnectivity, potentially negating the benefits of modularization.

To ensure the robustness of the proposed decomposition for this model, the aim is to employ a multi-objective evaluation criterion that simultaneously optimizes for minimized inter-service coupling, maximized intra-service cohesion, and adherence to defined service size constraints.

We completed this model by incorporating the modularity quality metric, which serves as a comprehensive indicator of

the partitioning effectiveness by calculating the ratio of internal interactions to total interactions, expressed as:

$$Q = \frac{W_{intra} - W_{inter}}{W_{total}}$$

where [see Eq. (14)],

$$W_{intra} = \sum_{(i,j)\in E} w_{ij} y_{ij}, \quad W_{inter} = \sum_{(i,j)\in E} w_{ij}(1 - y_{ij}),$$

$$W_{total} = \sum_{(i,j)\in E} w_{ij} \tag{14}$$

This metric evaluates the structural cohesiveness of classes within a partition against the coupling between partitions, where a higher ratio indicates superior partition quality by ensuring that intra connectivity significantly outweighs the inter-connectivity between identified microservices.

From all the parameters defined above, we delivered a multi-object optimization as:

$$\min_{(i,j)\in E} w_{ij}(1 - y_{ij}), \quad \max_{j} Cohesion_j$$

This is subject to [see Eq. (15)]:

$$L_{\min} \leq \sum_{i} x_{i,j} \leq L_{\max} \tag{15}$$

The unified objective function is derived as follows [see Eq. (16)]:

$$\max \left[\lambda_1 \frac{W_{intra} - W_{inter}}{W_{total}} \right.$$

$$\left. + \lambda_2 \sum_{j} \frac{|E_j^{int}|}{|V_j|} - \lambda_3 \sum_{(i,j)\in E} w_{ij}(1 - y_{ij}) \right] \tag{16}$$

This final model utilizes quality metrics such as coupling, cohesion, complexity, and the size employed to determine the optimal clustering output and maintainability quality, where higher scores are assigned to configurations demonstrating high cohesion and low coupling.

D. Optimal Microservice Partitioning Algorithm

The optimal microservice partitioning algorithm is presented in Algorithm 1.

Algorithm 1 Optimal Microservice Partitioning

Require: Class set $C_M = \{c_1, c_2, \ldots, c_N\}$, dependency graph $G(V, E)$
Require: Weight parameters $\alpha, \omega_{ec}, \omega_{ac}, \lambda_1, \lambda_2$
Require: Constraints $S_{\min}, S_{\max}, L_{\min}, L_{\max}$
Require: Static metrics: cyclomatic complexity $V(G)_c$, method counts MC_c
Require: Commit history for co-change extraction
Require: Initial empty partition matrix $X = [x_{i,j}]$
Require: Feasible initialization of service sizes
$X \leftarrow$ random feasible assignment of classes to microservices
$M^* \leftarrow \emptyset$
$t \leftarrow 0$
Compute initial coupling weights w_{ij}
while convergence criterion is not satisfied **do**
  $t \leftarrow t + 1$ ▷ Update iteration counter
  Compute structural metrics $S_i,\ V(G)_c,\ MC_c$
  Compute scalability metric L_i
  Compute cohesion for each service $Cohesion_j$
  Update coupling weights w_{ij}
  Evaluate modularity Q
  Evaluate objective function $F(M)$
  Update class assignments $x_{i,j}$ using optimization operator
  (Non dominated Sorting Genetic Algorithm II (NSGA-II)
  Enforce constraints:
  $S_{\min} \leq S_i \leq S_{\max}$
  $L_{\min} \leq |m_j| \leq L_{\max}$
  $\sum_{j=1}^{K} x_{i,j} = 1, \quad \forall c_i \in C_M$
  if $F(M) > F(M^*)$ **then**
    $M^* \leftarrow M$
    $F(M^*) \leftarrow F(M)$
  end if
end while

E. Optimization Configuration

The unified objective function was optimized using the Non-dominated Sorting Genetic Algorithm II (NSGA-II)[25], a widely adopted evolutionary algorithm for multi-objective optimization problems.

The optimization process employed the following parameter settings: Population size: 100 candidate decompositions, Number of generations: 500, Crossover probability: 0.85, Mutation probability: 0.10, Tournament size: 2 Elitism strategy: Enabled

Each candidate solution was encoded as a class-to-service assignment vector. The fitness function was computed according to Eq. (16), balancing modularity, cohesion, and coupling objectives. The optimization process terminated when either:

- The maximum number of generations was reached, or
- The improvement in objective value remained below 0.001 for 50 consecutive generations.

This convergence criterion ensured computational efficiency while maintaining solution quality.

F. Integration with Domain-Driven Design (DDD)

The model we presented incorporates Domain-Driven Design concepts, specifically the definition of bounded contexts, to ensure that the refactoring process aligns the technical boundaries of microservices with the distinct business domains they represent. By leveraging these DDD principles, the model resolves the inherent ambiguity in defining bounded context boundaries by supplementing logical domain divisions with quantitative coupling and cohesion metrics to ensure high modularity and optimized service granularity [8]. This proposed model complements Domain-Driven Design by transforming domain concepts into quantifiable optimization criteria.

Initially, DDD is used to define candidate bounded contexts; these candidates are then refined using the optimization algorithm, where each m_i corresponds to a hypothesized bounded context derived from domain analysis:

$$M^{(0)} = \{m_1, m_2, \ldots, m_K\}$$

G. Mapping Between DDD and Model

Table I presents the mapping between DDD concepts and mathematical model components.

TABLE I. MAPPING BETWEEN DDD CONCEPTS AND MATHEMATICAL MODEL COMPONENTS

DDD Concept	Mathematical Representation
Bounded Context	Microservice m_i
Domain Cohesion	Size metric S_i
Service Autonomy	Scalability Sc_i
Context Independence	Low Coupling C_i^{ext}

V. CASE STUDY: IOT IRRIGATION SYSTEM

The IoT irrigation system was specifically chosen as the subject of this investigation due to its complex, multi-domain architecture and the distinct scalability challenges inherent in its sensor data processing and control mechanisms [28], which necessitate a rigorous decomposition strategy to validate the proposed mathematical model.

This system, deployed and located in Bugesera District in Rwanda, is characterized by a monolithic structure where sensor data acquisition, irrigation control logic, user interface management, and data persistence are tightly coupled, making it an ideal candidate for demonstrating how Domain-Driven Design can identify bounded contexts for separation into independent microservices. The selection of an IoT irrigation system as the subject of this case study provides a relevant context for validating the mathematical model, as such systems typically require high scalability to manage fluctuating sensor data loads and distinct service boundaries initially deployed as a single deployable unit for the irrigation system.

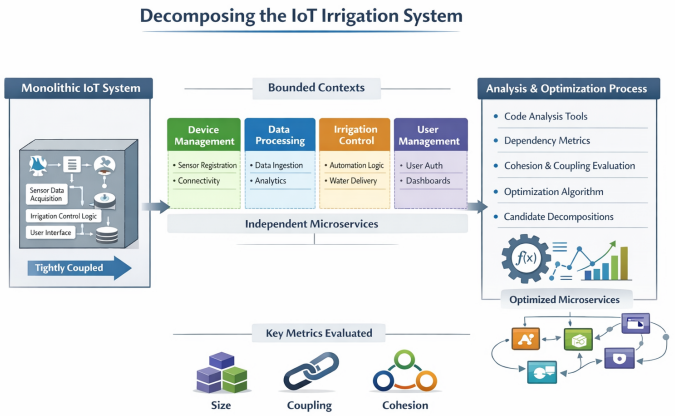


Fig. 2. System analysis and optimization framework.

We analyzed the code base and system requirements of the IoT irrigation system implemented in Bugesera District. As illustrated in Fig. 2, the analysis reveals specific functional domains to define the bounded contexts necessary for the refactoring process. The functional decomposition identified four primary bounded contexts: *Device Management*, which oversees sensor registration and connectivity; *Data Processing*, responsible for ingesting and analyzing voluminous sensor readings; *Irrigation Control*, encompassing the automation logic for water delivery based on sensor inputs; and *User Management*, handling authentication, user profiles, and dashboard interactions.

To quantitatively assess the performance of the proposed mathematical model, empirical data were collected from the existing monolithic IoT irrigation system by analyzing source code dependencies, class interaction frequencies, and module complexity metrics. These measurements established baseline values for service size, coupling, and scalability potential.

A. Data Collection, Analysis, and Refactoring Process of IoT Irrigation System

The data collection process involved mapping the static code metrics against the four identified bounded contexts to calculate specific coefficients for size, coupling, and scalability within the mathematical model.

These coefficients serve as the foundational parameters for the optimization algorithm, determining the weight of each quality attribute in the objective function designed to drive the automated partitioning of the monolith into microservices. Consequently, these coefficients are processed by the optimization algorithm to evaluate candidate decompositions, wherein solutions that minimize coupling while maximizing internal cohesion are prioritized to identify the optimal service boundaries, as demonstrated in [7].

This granular mapping of functionality and dependencies serves as the empirical foundation for generating multiple candidate decompositions, which are then rigorously evaluated against quality metrics such as cohesion and coupling to select the most effective microservice architecture [29].

To systematically gather the necessary data for modeling the IoT irrigation system, static analysis tools were employed to extract quantitative metrics regarding class dependencies, method invocations, and communication patterns between modules within the monolithic codebase. The dynamic analysis involved the extracted metrics, including different coupling values to measure inter-module dependencies, along with relational cohesion metrics to assess the functional relatedness of classes grouped within the same bounded context, thereby providing a multidimensional view of the system's structural characteristics.

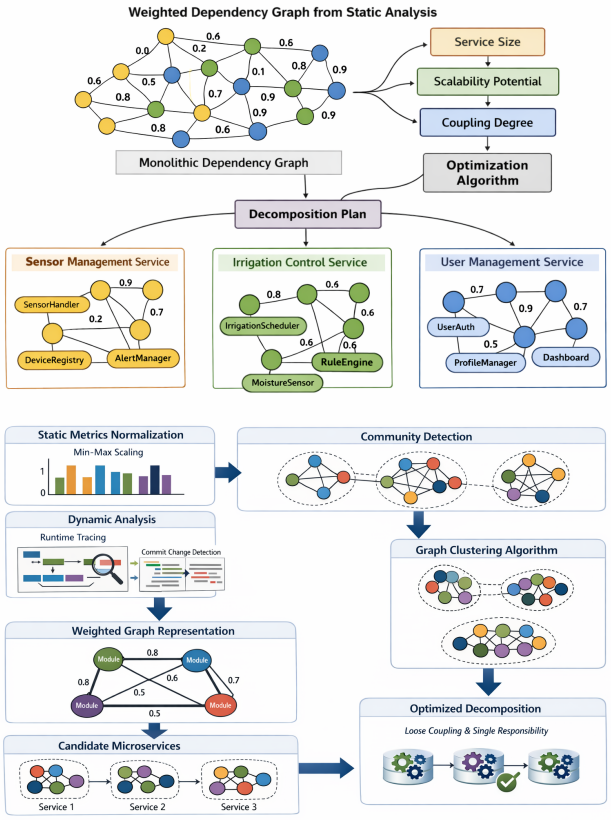


Fig. 3. Static analysis of IoT irrigation system (left) and decomposition criteria of the IoT irrigation system (right).

Moreover, graph clustering approaches, including community detection algorithms, are applied to detect highly cohesive clusters that correspond to potential microservices. Finally, combinatorial optimization techniques are utilized to ensure that the resulting decomposition satisfies key architectural principles, particularly loose coupling and single responsibility.

From static code analysis, the construction of a weighted dependency graph was derived (Fig. 3), where nodes represent classes and edges reflect the strength of relationships. This enables the algorithm to partition the system into highly cohesive communities corresponding to functional domains such as sensor management, scheduling logic, and user access control [8]. Once the dependency graph was established, the mathematical model calculated specific metrics for service size, scalability potential, and coupling degrees to quantitatively assess each identified functional domain.

This quantitative analysis utilized the calculated metrics as input variables for a multi-objective optimization algorithm to partition the dependency graph, identifying specific class clusters that minimize inter-service coupling while maintaining size and scalability constraints necessary for effective IoT deployment [18].

The execution of this decomposition plan required the strategic implementation of a database-per-service pattern and the definition of standardized communication APIs to ensure loose coupling and independent deployability. Specifically, the refactoring process involved isolating the sensor data ingestion logic into a dedicated service capable of independently scaling to handle high-frequency telemetry streams, while the irrigation control logic was encapsulated within a separate scheduling service to ensure that critical timing operations remain isolated from UI-induced latency.

Simultaneously, user management functions were extracted into an isolated authentication service to decouple access control mechanisms from core irrigation operations, ensuring that security updates do not disrupt the system's real-time processing capabilities.

Following the structural decomposition, the functional performance of the extracted microservices was assessed to ensure that the separation did not negatively impact the system's ability to process real-time telemetry and execute irrigation commands.

B. Evaluation of IoT Irrigation System

To process the evaluation, two architectures were deployed and observed to compare the results of latency, throughput, CPU usage, and the failure rate.

1) Architecture deployment: The data flow in the pipeline follows a sequence where data from all sensors passes through the ESP32. The presented data flow pipeline illustrates the end-to-end operation of an IoT-based irrigation system, where environmental data is continuously monitored, processed, and used to automate irrigation decisions. Each component contributes to a distributed and scalable architecture aligned with modern microservice and IoT design principles.

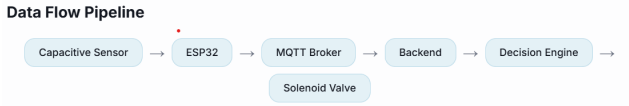


Fig. 4. Data flow pipeline

The process presented in Fig. 4 begins with the capacitive sensor that generates analog signals proportional to soil moisture levels. The ESP32 microcontroller acts as the edge device responsible for reading sensor data, converting analog signals to digital values, and performing lightweight preprocessing. In this architecture, the ESP32 enables real-time transmission of sensor data to the cloud infrastructure via wireless communication.

The MQTT broker serves as the central communication hub, implementing a publish-subscribe model. The backend system performs data storage, aggregation, and integration

with other services. The decision engine applies rules or machine learning models to determine irrigation actions. In this architecture, the decision logic is defined as:

```
if (soil_moisture < 30 && temperature > 25) {
    irrigation = true; // Turn ON irrigation
} else {
    irrigation = false; // Turn OFF irrigation
}
```

The solenoid valve acts as the physical actuator that controls water flow, opening to allow irrigation and closing to stop it.

After defining the data flow, the two architectures were deployed for comparison [see Fig. 5].

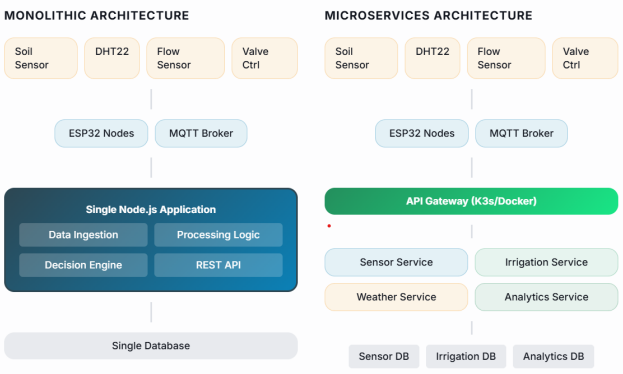


Fig. 5. Deployment of monolithic and microservice architecture.

C. System Analysis

To evaluate the efficiency gains derived from the architectural transition, comparative execution time measurements were conducted between the original monolithic application and the refactored microservice implementation. The results demonstrated that the microservice architecture achieved an increase of 1500 req/s compared to 800 req/s in the monolithic system.

TABLE II. PERFORMANCE COMPARISON

Metric	Monolithic	Microservices
Avg Latency	140 ms	180 ms
Throughput	800 req/s	1500 req/s
CPU Usage	85%	65%
Memory Usage	900 MB	1200 MB
Failure Impact	Total	Partial
Deployment	Simple	Complex
Scalability	Limited	Excellent
Dev Complexity	Low	High

Furthermore, response time analysis showed a decrease in latency under concurrent access, as the distributed architecture mitigated bottlenecks associated with the monolithic design. Load testing demonstrated stable response times with error rates below 0.1% at peak loads, indicating improved reliability due to isolated failure domains [see Fig. 6].

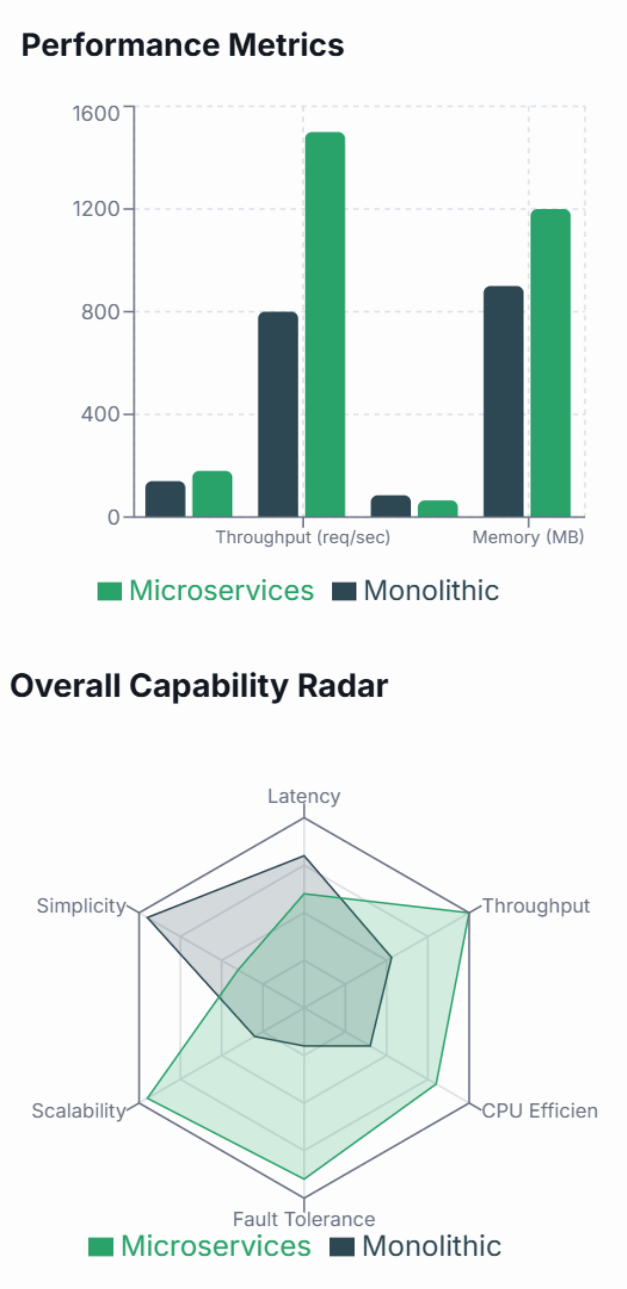


Fig. 6. Performance comparison (left) and overall capacity radar (right).

D. Scalability Analysis

System scalability was assessed by increasing the number of simulated IoT devices across four levels: 100, 1,000, 5,000, and 10,000 devices. At each level, latency was measured as the time between sensor data transmission and backend response.

TABLE III. LATENCY VS. NUMBER OF DEVICES

Devices	Monolithic Latency	Microservices Latency
100	120 ms	150 ms
1,000	200 ms	170 ms
5,000	350 ms	190 ms
10,000	Failure	230 ms

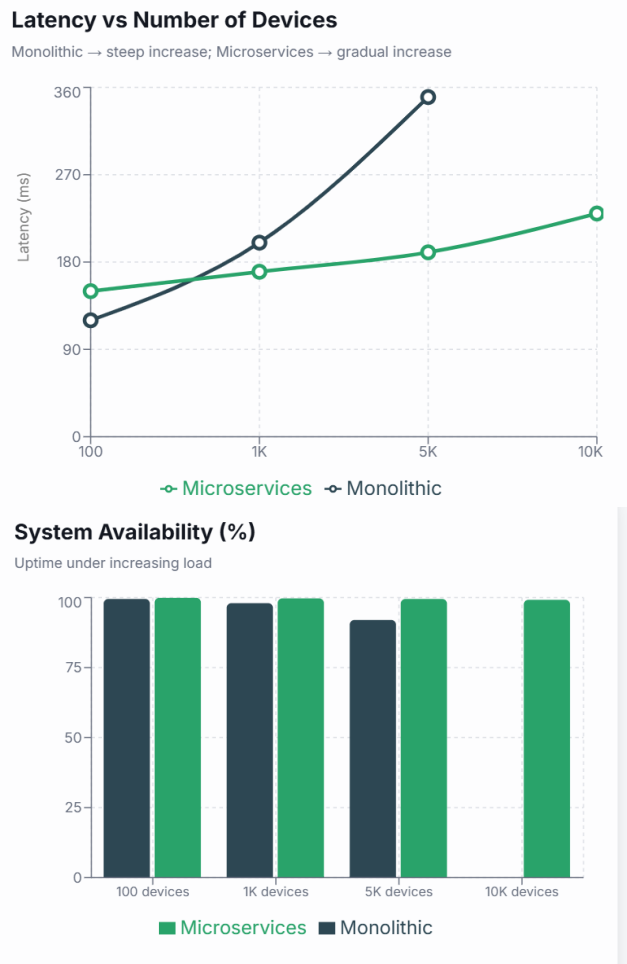


Fig. 7. Latency vs. devices (left) and system availability (right).

The results demonstrate that while monolithic architectures perform better at low load, their performance degrades significantly as system scale increases. In contrast, microservices maintain stable latency and operational reliability under high load conditions (see Fig. 7).

These findings align with the proposed mathematical model, where scalability optimization (Sc_i) and reduced coupling (C_i) contribute to improved system resilience and performance.

Additionally, structural improvements resulted in reduced coupling due to strict enforcement of bounded contexts and asynchronous communication. The refactored architecture achieved better resource utilization, with reduced CPU consumption and controlled memory usage during peak loads. This validates the effectiveness of size constraints in preventing resource contention and ensuring efficient IoT deployment.

Furthermore, scalability validation showed that the sensor ingestion service handled a 50% increase in throughput without proportional latency growth, confirming the model's effectiveness in predicting workload scaling behavior in dynamic IoT environments.

1) Sensitivity analysis: To evaluate the robustness of the proposed model, a sensitivity analysis was conducted by varying the weighting coefficients λ_1, λ_2, and λ_3 within the range [0.1,0.9].

The results demonstrated that the proposed decomposition remained stable across a wide range of parameter settings, with variations in modularity scores remaining below 5%. This indicates that the model is not excessively sensitive to small changes in optimization weights and therefore exhibits strong robustness for practical deployment.

E. Model Validation Results and Discussion

This section presents a comprehensive analysis of the empirical data derived from the case study, examining the efficacy of the mathematical model in optimizing the refactoring process and validating the theoretical assertions regarding service size, scalability, and coupling reduction.

The empirical validation of the mathematical model was conducted by comparing the predicted service boundaries against the actual operational metrics of the deployed microservices, confirming that the algorithm accurately identified optimal split points within the irrigation system's domain model. Furthermore, the model's scalability predictions were validated by subjecting the system to incremental load increases, where the projected resource allocation closely matched the actual consumption patterns of the sensor data and control services across varying operational intensities. This precise matching of resource utilization confirms that the mathematical framework effectively accommodates the dynamic scaling requirements inherent to IoT environments, thereby validating the model's predictive accuracy for deployment planning.

The quantification of interaction dependencies indicates that the mathematical model successfully translated Domain-Driven Design principles into executable decoupling strategies, yielding an architectural state that mirrors theoretical ideals where coupling scores remain significantly below acceptable thresholds.

To validate the effectiveness of the proposed unified optimization model, the empirical results obtained from the IoT irrigation system were analyzed using performance metrics illustrated in the radar chart. The validation focuses on assessing whether the architectural behavior observed in practice aligns with the theoretical expectations derived from the optimization objective function:

$$\max \left[\lambda_1 \frac{W_{intra} - W_{inter}}{W_{total}} \right]$$

$$+\lambda_2 \sum_j \frac{|E_j^{int}|}{|V_j|} - \lambda_3 \sum_{(i,j)\in E} w_{ij}(1 - y_{ij})]$$

1) Validation of scalability component: The experimental results demonstrate that the microservices architecture achieves significantly higher values in throughput, CPU efficiency, and scalability compared to the monolithic architecture. The improved throughput and CPU efficiency indicate that the ratio $\frac{L_i}{U_i}$ is maximized in the microservices architecture, meaning that more workload is processed per unit of resource consumption. Furthermore, the stable scalability observed under increasing load conditions confirms that the distributed nature

of microservices effectively preserves performance, aligning with the maximization objective of Sc_i.

F. Validation of Coupling Minimization

The results show a clear advantage of microservices in terms of fault tolerance, which serves as an indirect validation of reduced coupling. According to the model:

$$C_i^{ext} = \sum_{c_p \in m_i} \sum_{c_q \notin m_i} w_{pq}$$

Lower coupling implies higher service independence and improved fault isolation. The superior fault tolerance observed in the microservices architecture confirms that minimizing C_i^{ext} leads to increased system resilience. In contrast, the monolithic architecture exhibits lower fault tolerance due to tightly coupled components, validating the negative impact of high coupling on system reliability.

The average coupling score across the refactored system was recorded at 20.4%, which satisfies the theoretical requirement of remaining below 30% and aligns with empirical observations from successful microservice decompositions.

G. Validation of Size Constraint

The size penalty ensures balanced service granularity. The reduced simplicity observed in the microservices architecture reflects the intentional decomposition into multiple services, which increases architectural complexity but avoids oversized components. This behavior confirms that the penalty function effectively discourages monolithic aggregation while promoting modular decomposition.

This operational isolation demonstrates how the mathematical model's emphasis on bounded contexts translates into enhanced deployment velocity and reduced maintenance overhead. The observed modularization improves the continuous integration pipeline and strengthens fault isolation capabilities, ensuring that failures remain localized within specific bounded contexts.

H. Validation of Latency Trade-offs

Although the proposed microservice architecture demonstrated significant improvements in throughput, fault tolerance, scalability, and resource utilization, the experimental results revealed a measurable latency increase under low-load conditions. As shown in Table II, the average response latency increased from 140 ms in the monolithic architecture to 180 ms in the microservice architecture. Similarly, Table III indicates that at a deployment scale of 100 IoT devices, latency increased from 120 ms to 150 ms after decomposition.

This behavior is an expected consequence of service decomposition. Unlike monolithic systems, microservices introduce network communication overhead due to remote procedure calls, message serialization, service discovery, and API gateway interactions. At small workloads, these additional communication costs outweigh the benefits of service independence, resulting in slightly higher response times.

However, as the workload increases, the advantages of independent scalability become dominant. At 1,000 devices and beyond, the microservice architecture consistently outperformed the monolithic architecture, ultimately maintaining operational stability at 10,000 devices where the monolithic implementation failed. Therefore, the proposed model introduces a latency-performance trade-off in low-load environments but achieves superior scalability and resilience under realistic large-scale IoT deployment conditions.

This limitation is acknowledged as an inherent characteristic of distributed architectures and represents a trade-off accepted by the optimization model in favor of long-term scalability and maintainability.

I. Statistical Validation

The optimization experiment was executed 30 independent times to account for the stochastic nature of the evolutionary search process.

The mean modularity score obtained by the proposed model was 0.84 with a standard deviation of 0.03, indicating high consistency across repeated executions.

Furthermore, a paired t-test comparing the proposed model against the Louvain baseline demonstrated statistically significant improvements in cohesion and coupling reduction at a significance level of $p < 0.05$.

J. Comparison with Existing Refactoring Approaches

To establish a verifiable benchmark, the proposed UMDO model was compared against three representative decomposition approaches frequently reported in the literature:

- K-Means Clustering
- Louvain Community Detection [30]
- Manual Domain-Driven Design Decomposition [31]

K-Means represents traditional partitioning approaches based on feature similarity, while Louvain represents graph-community detection methods widely used in software modularization. Manual Domain-Driven Design decomposition represents expert-driven bounded-context identification without optimization support.

The comparison demonstrates that the proposed UMDO framework achieves the lowest coupling and highest cohesion and modularity scores among all evaluated approaches. These results indicate that integrating bounded-context knowledge with multi-objective optimization provides a more effective decomposition strategy than purely structural or manually derived approaches (see Table IV).

Existing microservice refactoring approaches rely on structural dependencies or semantic similarity techniques. While effective for modularity detection, these methods often neglect runtime and scalability considerations. The proposed model introduces a unified multi-objective optimization framework integrating service size, scalability, and coupling constraints, reducing subjectivity and improving decomposition accuracy.

TABLE IV. QUALITATIVE AND QUANTITATIVE COMPARISON OF THE PROPOSED UMDO MODEL WITH EXISTING MICROSERVICE DECOMPOSITION APPROACHES.

Criterion	K-Means Clustering	Louvain Community Detection	DDD Manual Decomposition	Proposed UMDO Model
Methodology	Structural clustering	Graph community detection	Domain-Driven Design	Multi-objective optimization
Coupling Handling	Implicit	Graph-based	Expert judgment	Explicit quantitative modeling
Scalability	Limited	Limited	Moderate	Runtime-aware modeling
Granularity Control	Heuristic	Community partitioning	Manual refinement	Formal size constraints
Clustering Strategy	Distance-based	Graph modularity	Business domains	Optimization-based refinement
Semantic Awareness	Limited	Limited	High	DDD-integrated validation
Bounded Context Identification	Not supported	Partially supported	Manual	Quantitative derivation
Adaptability	General-purpose	General-purpose	Application-dependent	IoT-aware scalability
Coupling Score	0.38	0.31	0.28	**0.20**
Cohesion Score	0.61	0.70	0.74	**0.86**
Modularity Score	0.59	0.68	0.72	**0.84**

VI. CONCLUSION

This study presents a mathematical framework for refactoring monolithic applications into microservices using Domain-Driven Design principles and quantitative metrics. The empirical results from the IoT irrigation case study confirm that the model effectively identifies optimal service boundaries by minimizing coupling and maximizing cohesion.

The findings demonstrate that integrating quantitative metrics with Domain-Driven Design provides a deterministic and objective approach to microservice decomposition, outperforming heuristic-based methods and improving scalability, maintainability, and system reliability. The experimental evaluation demonstrated statistically significant improvements over graph clustering and manual decomposition baselines, while sensitivity analysis confirmed the robustness of the optimization framework across varying parameter configurations.

VII. FUTURE WORK

Future research could extend this work by integrating cloud-native and AI-based optimization techniques to further enhance scalability. Additionally, incorporating dynamic runtime metrics could enable adaptive microservice refactoring in real-time, particularly for IoT environments with fluctuating workloads.

Another promising direction is the application of transfer learning to improve model generalization across different technology stacks. Finally, establishing standardized benchmarks and open datasets for microservice partitioning would facilitate comparative analysis and broader validation of decomposition models.

REFERENCES

[1] I. Trabelsi, B. Mahmoudi, J. B. Minani, N. Moha, and Y. G. Gueheneuc, "A systematic literature review of machine learning approaches for migrating monolithic systems to microservices," in *IEEE Transactions on Software Engineering*, vol. 51. Institute of Electrical and Electronics Engineers Inc., 11 2025, pp. 2972–2995.

[2] A. Chansarkar, "Strategic development patterns for cloud-native enterprise solutions: Balancing optimization, scalability, and architecture," *Journal of Information Systems Engineering and Management*, vol. 2025, pp. 2468–4376, 2025. [Online]. Available: https://www.jisem-journal.com/

[3] V. L. Nogueira, F. S. Felizardo, A. M. Amaral, W. K. Assuncao, and T. E. Colanzi, "Insights on microservice architecture through the eyes of industry practitioners," in *Proceedings - 2024 IEEE International Conference on Software Maintenance and Evolution, ICSME 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 765–777.

[4] M. G. Amri, T. Raharjo, A. N. Fitriani, N. Rasyid, and P. Hutasuhut, "Critical success factors of microservices architecture implementation in the information system project," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 15, 2024. [Online]. Available: www.ijacsa.thesai.org

[5] M. I. Joselyne, G. Bajpai, and F. Nzanywayingoma, "A systematic framework of application modernization to microservice based architecture," in *7th International Conference on Engineering and Emerging Technologies, ICEET 2021*. Institute of Electrical and Electronics Engineers Inc., 2021.

[6] N. A. Mansour, H. Sbai, and K. Baina, "Minds: An nlp-driven process mining framework for superior microservices decomposition," *IEEE Access*, vol. 13, pp. 208 376–208 396, 2025.

[7] G. Filippone, N. Q. Mehmood, M. Autili, F. Rossi, and M. Tivoli, "From monolithic to microservice architecture: an automated approach based on graph clustering and combinatorial optimization," in *Proceedings - IEEE 20th International Conference on Software Architecture, ICSA 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 47–57.

[8] H. Vural and M. Koyuncu, "Does domain-driven design lead to finding the optimal modularity of a microservice?" *IEEE Access*, vol. 9, pp. 32 721–32 733, 1 2021.

[9] T. Rathod, C. T. Joseph, and J. P. Martin, "Improving industry 4.0 readiness: Monolith application refactoring using graph attention networks," in *Proceedings - 23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing Workshops, CCGridW 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 223–230.

[10] I. A. Kautsar, M. R. Maika, A. N. Budiman, A. B. Setyawan, and J. Y. Awali, "Microservice based architecture: The development of rapid prototyping supportive tools for project based learning," in *EDUNINE 2023 - 7th IEEE World Engineering Education Conference: Reimaging Engineering - Toward the Next Generation of Engineering Education, Merging Technologies in a Connected World, Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2023.

[11] E. Volynsky, M. Mehmed, and S. Krusche, "Architect: A framework for the migration to microservices," in *Proceedings - 2022 International Conference on Computing, Electronics and Communications Engineering. iCCECE 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 71–76.

[12] C. Y. Li, S. P. Ma, and T. W. Lu, "Microservice migration using strangler fig pattern: A case study on the green button system," in *Proceedings - 2020 International Computer Symposium, ICS 2020*. Institute of Electrical and Electronics Engineers Inc., 12 2020, pp. 519–524.

[13] M. Bhandari, "Microservices vs. monolithic architectures in real-time distributed systems: A comparative analysis," *Journal of Information Systems Engineering and Management*, vol. 2025, pp. 2468–4376, 2025. [Online]. Available: https://www.jisem-journal.com/

[14] D. N. Reddy, R. Suryodai, V. Kumar, M. Ambika, E. Muniyandy, V. R. Krishna, and B. Abdurasul, "A scalable microservices architecture for real-time data processing in cloud-based applications," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 16, p. 2025, 2025. [Online]. Available: www.ijacsa.thesai.org

[15] H. Hassan, M. A. Abdel-Fattah, and W. Mohamed, "Migrating from monolithic to microservice architectures: A systematic literature review," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 15, 2024. [Online]. Available: www.ijacsa.thesai.org

[16] N. Legowo, Erin, E. H. Lee, and M. Djakaria, "Designing service oriented architecture model in sehatin application with a domain driven design approach," in *Proceedings of 2023 International Conference on Information Management and Technology, ICIMTech 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 678–683.

[17] S. Weerasinghe and I. Perera, "Optimized strategy for inter-service communication in microservices," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, 2023. [Online]. Available: www.ijacsa.thesai.org

[18] B. Liu, J. Xiong, Q. Ren, S. Tyszberowicz, and Z. Yang, "Log2ms: A framework for automated refactoring monolith into microservices using execution logs," in *Proceedings - IEEE International Conference on Web Services, ICWS 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 391–396.

[19] G. Baldoni, J. Quevedo, C. Guimaraes, A. D. L. Oliva, and A. Corsaro, "Data-centric service-based architecture for edge-native 6g network," *IEEE Communications Magazine*, vol. 62, pp. 32–38, 4 2024.

[20] V. Goshika, "Demystifying distributed microservices architecture for enterprise-scale systems," *International Journal of Computational and Experimental Science and Engineering*, vol. 12, 2 2026. [Online]. Available: https://www.ijcesen.com/index.php/ijcesen/article/view/4860

[21] H. Farsi, D. Allaki, A. En-Nouaary, and M. Dahchour, "A graph-based solution to deal with cyclic dependencies in microservices architecture," in *Proceedings - 2022 International Conference on Future Internet of Things and Cloud, FiCloud 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 254–259.

[22] M. Raji, A. Hota, T. Hobson, and J. Huang, "Scientific visualization as a microservice," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 1760–1774, 2020.

[23] U. K. Chilakalapalli, B. Mohan, and V. R. Surasani, "From legacy to cloud: Migration strategies for traditional financial institutions using aws," Tech. Rep., 2025. [Online]. Available: www.ijacsa.thesai.org

[24] V. Kumar, "Microservice-driven performance optimization in large-scale transaction processing systems," *International Journal of Computational and Experimental Science and Engineering*, vol. 11, 12 2025.

[25] V. Cortellessa, D. D. Pompeo, V. Stoico, and M. Tucci, "On the impact of performance antipatterns in multi-objective software model refactoring optimization," in *Proceedings - 2021 47th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2021*. Institute of Electrical and Electronics Engineers Inc., 9 2021, pp. 224–233.

[26] D. A. D'Aragona, L. Pascarella, A. Janes, V. Lenarduzzi, and D. Taibi, "Microservice logical coupling: A preliminary validation," in *Proceedings - IEEE 20th International Conference on Software Architecture Companion, ICSA-C 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 81–85.

[27] O. Al-Debagy and P. Martinek, "Extracting microservices' candidates from monolithic applications: Interface analysis and evaluation metrics approach," in *SOSE 2020 - IEEE 15th International Conference of System of Systems Engineering, Proceedings*, 2020, pp. 289–293.

[28] I. J. Munezero, F. Mukamanze, H. Eric, A. Ngenzi, J. O. Sinayobye, T. Murangira, C. Uwera, P. Muvunyi, and P. Ikundabayo, "A microservice based iot framework with machine learning approaches for drip irrigation scheduling in rwanda's cyohoha sud region," Tech. Rep., 2024. [Online]. Available: https://www.jisem-journal.com/

[29] T. Lopes and A. R. Silva, "Monolith microservices identification: Towards an extensible multiple strategy tool," in *Proceedings - IEEE 20th International Conference on Software Architecture Companion, ICSA-C 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 111–115.

[30] E. Gaidels and M. Kirikova, "Service dependency graph analysis in microservice architecture," in *Perspectives in Business Informatics Research*, R. A. Buchmann, A. Polini, B. Johansson, and D. Karagiannis, Eds. Cham: Springer International Publishing, 2020, pp. 128–139.

[31] J. Ocharan-Hernandez, J. O. Limon, X. Cortes-Verdin, and M. K. Sangabriel-Alarcon, "Domain-driven design in microservices-based systems development: A systematic literature review and thematic analysis," *Programming and Computer Software*, vol. 50, pp. 742–770, 2024.

A Multi-Criteria Decision-Making Model for ERP Selection in Moroccan SMEs Using Fuzzy AHP

Yassine Zouhair, Younous Elmrini
SIAD Laboratory, Hassan First University, Settat, Morocco

Abstract—Today, selecting an Enterprise Resource Planning (ERP) system has become a strategic decision for Small and Medium-sized Enterprises (SMEs) seeking to improve their operational efficiency, modernize their information systems, and strengthen their competitiveness in an increasingly digitalized environment. In this context, identifying the most relevant ERP selection criteria is essential to support decision-makers in choosing the solution that best fits their organizational and strategic needs. However, previous studies have generally examined ERP selection criteria from a broad perspective, without prioritizing their relative importance according to the specific context of SMEs. The objective of this study is therefore to identify and prioritize the ERP selection criteria that are most critical for Moroccan SMEs, as well as to propose a decision-making model that supports ERP evaluation and selection. To achieve this objective, a mixed-methods approach was adopted. Semi-structured interviews with ERP professionals and practitioners were conducted to identify the most relevant criteria in the Moroccan SME context. In addition, the Fuzzy Analytic Hierarchy Process (Fuzzy AHP) method was used to evaluate and weight these criteria under conditions of uncertainty and subjective judgment. The results reveal that functional scope is the most critical criterion in ERP selection, followed by ease of use and total cost. Technical criteria such as compatibility and security are considered moderately important, whereas scalability appears to have less influence on the selection decision. Based on these findings, a decision-making model is proposed to help SMEs evaluate ERP alternatives more effectively and reduce the risks associated with ERP selection. This research contributes to the existing ERP literature by providing a structured and prioritized analysis of ERP selection criteria specifically adapted to Moroccan SMEs. It also offers ERP consultants and decision-makers a practical framework for assessing ERP solutions and improving the quality and reliability of ERP selection decisions.

Keywords—ERP selection; Fuzzy AHP; Moroccan SMEs; multi-criteria decision-making; information systems; decision framework; digital transformation

I. Introduction

In an ERP, the business processes related to procurement, sales, finance, inventory, and customer relationships can be managed in an efficient way by centralizing and integrating them into one information system. However, implementing an ERP within SMEs represents a significant investment and requires time before it can be fully integrated into the business.

In order to improve their productivity, increase their efficiency, and make themselves better prepared to tackle the challenges in the highly competitive environment, some companies, including Moroccan SMEs, have opted for ERP adoption within their IS.

Consequently, choosing an ERP should not be a hasty decision. It requires a very thorough analysis of several factors, including organizational structure, business processes, business objectives, and user expectations. Given the wide variety of ERP solutions available, selecting the most suitable ERP is a major strategic decision [1].

This decision depends largely on the definition of selection criteria and their prioritization. Indeed, the selected criteria have a direct influence on the success of the ERP implementation within the company. Clearly ranking these criteria by priority helps effectively guide the selection process and accelerate decision-making.

Given the high costs associated with ERP implementation, it is essential to ensure that the chosen solution effectively meets the company's needs [2]. Companies that succeed in this choice are often those that have a strong grasp of their business processes as well as the needs of their employees.

Research has been conducted to assist companies in their selection process [3,4,5,6,7]. However, there is a lack of research aimed at identifying the criteria for selecting an ERP as well as a decision-making model for ERP selection, particularly in the context of Moroccan SMEs.

In this context, this study aims to identify and prioritize ERP selection criteria, as well as to propose a decision-making model for ERP selection in Moroccan SMEs.

II. Literature Review

Numerous studies have proposed decision-making models and methods—particularly those based on multi-criteria approaches—to structure the ERP selection process. Table I provides an overview of the key studies identified in the literature regarding the selection of ERP.

After reviewing the literature, we found that there is no multi-criteria decision-making model for selecting an ERP within Moroccan SMEs. The studies were conducted in contexts different from that of Moroccan SMEs, which limits their applicability to this context.

TABLE I. SUMMARY OF STUDIES ON ERP SÉLECTION

Authors	Country	Abstract
[3]	United Arab Emirates	Identifies and ranks the critical factors for ERP selection in UAE ports using FDM and RII, highlighting cybersecurity and stakeholder integration as priorities.
[8]	Turkey	Proposes a hybrid method combining hesitant fuzzy cross-entropy and probabilistic hesitant fuzzy COPRAS for ERP selection, improving the ranking of alternatives based on multiple criteria.
[9]	Indonesia	Proposes a hybrid Fuzzy AHP–TOPSIS method for selecting an open-source ERP for SMEs in Indonesia, emphasizing system quality (security, reliability) as a key criterion.
[4]	India	Uses Fuzzy AHP to select ERP systems and ERP consultants. Identifies the following as key criteria: reputation, ERP experience, and cost.
[5]	India	Analyzes the factors influencing ERP selection using Fuzzy AHP. Cost is the primary criterion, followed by usability and business requirements.
[10]	Portugal	Develops a hybrid MCDM model combining MMASSI/IT and AHP to evaluate and select an ERP, validated in an industrial setting.
[6]	Turkey	Proposes a multi-criteria decision-making (MCDM) approach that combines fuzzy ANP and GRA to select an ERP while accounting for uncertainty, thereby improving the ranking of alternatives.
[11]	India	Proposes a framework that incorporates critical success factors (CSFs) into ERP selection and introduces a Fuzzy TOPSIS approach to improve decision-making.

III. RESEARCH QUESTIONS

Given this gap in the literature, this study aims to answer the following question:

What are the most relevant criteria for selecting an ERP in Moroccan SMEs, and how can these criteria be organized to develop a decision-making model for ERP selection tailored to their specific context?

We will attempt to answer this central question through the following sub-questions:

1) What are the key criteria to consider when selecting an ERP for Moroccan SMEs?

2) Which criteria have the greatest influence on the selection of an ERP for Moroccan SMEs?

3) How can we develop a decision-making model for ERP selection that is relevant, practical, and tailored to Moroccan SMEs?

IV. RESEARCH METHODOLOGY

To answer these research questions, we adopted a mixed research method.

A. Qualitative Research Method

To answer the first question, we adopted a qualitative research method to identify the criteria for selecting an ERP within Moroccan SMEs. To this effect, semi-structured interviews were conducted among employees working for SMEs of Morocco, including IT managers, financial specialists, and key users, alongside experts. Such an interviewing technique would enable us to adopt both functional and specialist views. The respondents have been carefully chosen based on a targeted sampling technique, where preference was given to those with past experience in ERP systems' usage and selection.

37 participants were interviewed for a period ranging from 30 to 45 minutes. The number of interviewees used was decided upon using the concept of theoretical saturation, meaning that additional interviews were stopped once they no longer yielded significantly new information.

Analysis of the obtained data helped determine the most prevalent and significant criteria. This enabled the creation of a structured list of ERP selection criteria within Moroccan SMEs, which was subsequently used in a quantitative analysis to evaluate and prioritize them.

B. Fuzzy Analytic Hierarchy Process

To answer the second question, we adopted a multi-criteria approach using the Fuzzy AHP.

AHP is a methodology for multi-criteria decision making. It allows structuring any complicated decision by hierarchically organizing decision criteria and assigning weights to them according to expert evaluations [12]. Expert evaluations could sometimes include some degree of vagueness. Therefore, a modification of AHP that includes fuzziness in the process was developed [13].

Data for this section of the research were collected from employees of SMEs operating in Morocco, including companies that have introduced an ERP solution and others who haven't yet used any similar software solutions. Participants were asked to make pairwise comparisons between the various identified criteria in order to assess their relative importance in the ERP selection process.

The method applied consists of five main steps: constructing the fuzzy comparison matrix, calculating fuzzy sums, determining synthetic values, defuzzifying weights, and final normalization.

1) Construction of the fuzzy comparison matrix

a) Define the criteria

List the criteria:

$$C_1, C_2, \ldots, C_N$$

b) Pairwise Comparison (Question for employees): The criteria are compared in pairs by employees in order to assess their relative importance in the ERP selection process.

For each pair C_i, C_j, participants are asked to indicate which criterion is more important and to what extent.

c) Conversion to a Triangular Fuzzy Number (TFN): Each response is converted to a Triangular Fuzzy Number:

$$\tilde{a}_{\{ij\}} = \left(l_{\{ij\}}, m_{\{ij\}}, u_{\{ij\}}\right)$$

Interpretation:

- l = lowest possible value
- m = most likely value

- u = highest possible value

Linguistic judgments are converted into fuzzy numbers according to the scale shown in Table II.

TABLE II. CONVERTING LINGUISTIC JUDGMENTS INTO TRIANGULAR FUZZY NUMBERS

Linguistic variables	Fuzzy scale	Corresponding fuzzy scale
Equally important	(1, 1, 1)	(1, 1, 1)
Moderately important	(1/2, 1, 3/2)	(2/3, 1, 2)
Important	(3/2, 2, 5/2)	(2/5, 1/2, 2/3)
Very important	(5/2, 3, 7/2)	(2/7, 1/3, 2/5)
Extremely important	(7/2, 4, 9/2)	(2/9, 1/4, 2/7)

d) Aggregation (average) when there are multiple consultants: When multiple employees participate in the evaluation, and each provides (l_k, m_k, u_k), then the final cell is:

$$l_{\{ij\}} = \frac{1}{K}\sum_{k=1}^{k} l_{i,j}^{(K)}$$

$$m_{\{ij\}} = \frac{1}{K}\sum_{k=1}^{k} m_{i,j}^{(K)}$$

$$u_{\{ij\}} = \frac{1}{K}\sum_{k=1}^{k} u_{i,j}^{(K)}$$

where, K represents the number of employees.

Therefore:

$$\tilde{a}_{\{ij\}} = \left(\frac{1}{K}\sum_{k=1}^{k} l_{i,j}^{(K)}, \frac{1}{K}\sum_{k=1}^{k} m_{i,j}^{(K)}, \frac{1}{K}\sum_{k=1}^{k} u_{i,j}^{(K)}\right)$$

e) Consistency rules for the matrix

Diagonal

A criterion compared to itself is equal to:

$$\tilde{a}_{\{ii\}} = (1,1,1)$$

Reciprocal (inverse)

If:

$$\tilde{a}_{\{ij\}} = (l, m, u)$$

So:

$$\tilde{a}_{\{ji\}} = \tilde{a}_{\{ij\}}^{\{-1\}} = \left(\frac{1}{u}, \frac{1}{m}, \frac{1}{l}\right)$$

f) Result of this step

We obtain the fuzzy comparison matrix:

$$\tilde{A} = \begin{bmatrix} 1,1,1 & \tilde{a}_{\{12\}} & \tilde{a}_{\{13\}} & \tilde{a}_{\{14\}} & \tilde{a}_{\{15\}} & \tilde{a}_{\{16\}} \\ \tilde{a}_{\{21\}} & 1,1,1 & \tilde{a}_{\{23\}} & \tilde{a}_{\{24\}} & \tilde{a}_{\{25\}} & \tilde{a}_{\{25\}} \\ \tilde{a}_{\{31\}} & \tilde{a}_{\{32\}} & 1,1,1 & \tilde{a}_{\{34\}} & \tilde{a}_{\{35\}} & \tilde{a}_{\{36\}} \\ \tilde{a}_{\{41\}} & \tilde{a}_{\{42\}} & \tilde{a}_{\{43\}} & 1,1,1 & \tilde{a}_{\{45\}} & \tilde{a}_{\{46\}} \\ \tilde{a}_{\{51\}} & \tilde{a}_{\{52\}} & \tilde{a}_{\{53\}} & \tilde{a}_{\{54\}} & 1,1,1 & \tilde{a}_{\{56\}} \\ \tilde{a}_{\{61\}} & \tilde{a}_{\{62\}} & \tilde{a}_{\{63\}} & \tilde{a}_{\{64\}} & \tilde{a}_{\{65\}} & 1,1,1 \end{bmatrix}$$

2) Calculation of fuzzy sums by criterion: For each criterion, the sum of the fuzzy values in the corresponding row is calculated:

$$\Sigma l_i = \sum_{j=1}^{n} l_{i,j}$$

$$\Sigma m_i = \sum_{j=1}^{n} m_{i,j}$$

$$\Sigma u_i = \sum_{j=1}^{n} u_{i,j}$$

These values represent the cumulative weight of each criterion relative to the others.

The total sums are then calculated:

$$\Sigma l_{tot} = \sum_{i=1}^{n} \Sigma l_i$$

$$\Sigma m_{tot} = \sum_{i=1}^{n} \Sigma m_i$$

$$\Sigma u_{tot} = \sum_{i=1}^{n} \Sigma u_i$$

3) Calculation of fuzzy synthetic values: Fuzzy synthetic values are obtained through normalization:

$$S_{\{l,i\}} = \Sigma l_i \times \frac{1}{\Sigma u_{tot}}$$

$$S_{\{m,i\}} = \Sigma m_i \times \frac{1}{\Sigma m_{tot}}$$

$$S_{\{u,i\}} = \Sigma u_i \times \frac{1}{\Sigma l_{tot}}$$

These values represent the relative importance of the criteria in a fuzzy manner.

4) Defuzzification of weights: To obtain a single usable weight, fuzzy values are converted into simple numerical values using the center-of-gravity method:

$$W_i = \frac{S_{\{l,i\}} + S_{\{m,i\}} + S_{\{u,i\}}}{3}$$

This step provides a numerical value representing the significance of each criterion.

5) Normalization of weights: The weights obtained are then normalized so that their sum equals 1:

$$w_i = \frac{W_i}{\sum_{k=1}^{n} W_k}$$

The weights can then be expressed as a percentage:

$$Weight_i(\%) = 100 * w_i$$

These values correspond to the final relative importance of each criterion.

V. ERP Selection Criteria

Data collected from 37 participants was analyzed to identify the most common and relevant criteria. Table III given below presents a structured list of ERP selection criteria within Moroccan SMEs.

TABLE III. List of ERP Selection Criteria Within Moroccan SMEs

Criteria code	Main criteria	Sub-criteria	Criteria code
C1	Functional scope of the ERP system	Accounting and Financial Management	C1.1
		Purchasing and Sales Management	C1.2
		Inventory and Logistics Management	C1.3
		Production and Project Management	C1.4
C2	System Usability	Intuitive user interface	C2.1
		Ease of learning the system	C2.2
		Ease of navigation	C2.3
		Availability of documentation and user support	C2.4
C3	Total Cost of the ERP System	Cost of ERP licenses	C3.1
		Implementation cost	C3.2
		Annual maintenance cost	C3.3
		User training cost	C3.4
C4	Compatibility and Technical Integration	Compatibility with existing systems	C4.1
		Ease of data migration	C4.2
		Integration with other software	C4.3
		Compatibility with hardware infrastructure	C4.4
C5	System Security and Reliability	Data Protection and Confidentiality	C5.1
		User Access Rights Management	C5.2
		Data Backup and Restoration	C5.3
		System Stability	C5.4
C6	Scalability of the ERP System	Ability to add new modules	C6.1
		Adaptability to business growth	C6.2
		Ability to customize the system	C6.3
		Availability of future functional enhancements	C6.4

VI. Findings: Prioritization of Criteria and Sub-Criteria for Selecting an ERP

In this study, we identified six main criteria for selecting ERP among Moroccan SMEs, and for each of these criteria, we defined four sub-criteria. We will begin by determining the priority of each criterion, and then the priority of each sub-criterion associated with each of them.

A. Prioritization of Criteria

For the purpose of decision-making using the Fuzzy AHP approach, the criteria were assigned the following codes: C1, C2, C3, C4, C5, and C6 (see Table IV). The criteria and their respective codes are detailed below:

C1: Functional scope of the ERP system

C2: System Usability

C3: Total Cost of the ERP System

C4: Compatibility and Technical Integration

C5: System Security and Reliability

C6: Scalability of the ERP System

TABLE IV. FUZZY PAIRWISE COMPARISON MATRIX OF CRITERIA

	C1			C2			C3			C4			C5			C6		
	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u	l	m	u
C1	1	1	1	1,129	1,227	1,325	1,38	1,5	1,62	1,774	1,929	2,083	2,258	2,455	2,651	3,105	3,375	3,645
C2	0,754	0,814	0,885	1	1	1	1,124	1,222	1,32	1,446	1,571	1,697	1,84	2	2,16	2,53	2,75	2,97
C3	0,613	0,667	0,72	0,753	0,818	0,884	1	1	1	1,183	1,286	1,389	1,505	1,636	1,767	2,07	2,25	2,43
C4	0,477	0,519	0,56	0,585	0,636	0,687	0,716	0,778	0,84	1	1	1	1,171	1,273	1,375	1,61	1,75	1,89
C5	0,375	0,407	0,44	0,46	0,5	0,54	0,562	0,611	0,66	0,723	0,786	0,849	1	1	1	1,265	1,375	1,485
C6	0,273	0,296	0,32	0,335	0,364	0,393	0,409	0,444	0,48	0,526	0,571	0,617	0,669	0,727	0,785	1	1	1

TABLE V. FUZZY VALUES FOR SYNTHETIC EXTENT AND RELATIVE IMPORTANCE WEIGHTS OF THE CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C1	10,646	11,486	12,324	0,022	0,024	0,025	0,234	0,276	0,308	0,27	27
C2	8,694	9,357	10,032	0,022	0,024	0,025	0,191	0,225	0,251	0,22	22
C3	7,124	7,657	8,19	0,022	0,024	0,025	0,157	0,184	0,205	0,18	18
C4	5,559	5,956	6,352	0,022	0,024	0,025	0,122	0,143	0,159	0,14	14
C5	4,385	4,679	4,974	0,022	0,024	0,025	0,096	0,112	0,124	0,11	11
C6	3,212	3,402	3,595	0,022	0,024	0,025	0,071	0,082	0,09	0,08	8
Sum	39,62	42,54	45,47								

The findings demonstrate that the criterion C1 (Functional scope of the ERP system) is ranked first with the weight of 27%. Such a finding shows how crucial the issue of the functional scope of the ERP system is; it should have the capability of handling all business activities within the organization. With such a functional scope of the ERP system, an integrative approach can be adopted to address the operations without adopting other solutions.

The second most important criterion is C2 (System Usability), which accounts for 22% of the total. It shows that usability is one of the key elements in adopting the system. A user-friendly ERP system will be less resistant to change and will increase efficiency as well.

Criterion C3 (Total Cost of the ERP System) ranks third, each at 18%. This finding demonstrates that the total cost of the system in the course of its installation and maintenance is an important criterion in selection, especially in view of the SME sector.

Criterion C4 (Compatibility and Technical Integration) comes next and accounts for 14% of the score, as the integration of the ERP system with other systems must be taken into account to ensure continuity.

Criterion C5 (System Security and Reliability) accounts for 11% of the total score. While it does not receive the same level of emphasis as the criteria mentioned above, it cannot be discounted for its value in ensuring the continuity of operations and safeguarding confidential data.

Finally, criterion C6 (Scalability of the ERP System) has the lowest weight at 8%. This implies that the scalability aspect of the ERP system is not highly valued for the short term but will be an important aspect in the long run (see Table V).

B. Prioritization of Sub-Criteria

1) Functional Scope of the ERP System: During the decision-making process, where the criterion "Functional scope of the ERP system" will be assessed through the fuzzy AHP method, the following sub-criteria codes will be applied: C1.1, C1.2, C1.3, and C1.4 (see Table VI). Listed below are the sub-criteria codes together with the sub-criteria themselves:

C1.1: Accounting and Financial Management

C1.2: Purchasing and Sales Management

C1.3: Inventory and Logistics Management

C1.4: Production and Project Management

TABLE VI. FUZZY PAIRWISE COMPARISON MATRIX OF SUB-CRITERIA FOR THE FUNCTIONAL SCOPE CRITERION

	C1.1			C1.2			C1.3			C1.4		
	l	m	u	l	m	u	l	m	u	l	m	u
C1.1	1	1	1	0,90	1,05	1,20	1,05	1,20	1,35	1,90	2,10	2,30
C1.2	0,83	0,95	1,11	1	1	1	1,05	1,15	1,30	1,70	1,90	2,10
C1.3	0,74	0,83	0,95	0,77	0,87	0,95	1	1	1	1,55	1,70	1,90
C1.4	0,43	0,48	0,53	0,48	0,53	0,59	0,53	0,59	0,65	1	1	1

TABLE VII. FUZZY WEIGHTS AND NORMALIZED PRIORITIES OF FUNCTIONAL SCOPE SUB-CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C1.1	4,85	5,35	5,85	0,052	0,057	0,062	0,252	0,305	0,363	0,31	31
C1.2	4,58	5	5,51	0,052	0,057	0,062	0,238	0,285	0,342	0,29	29
C1.3	4,06	4,4	4,8	0,052	0,057	0,062	0,211	0,251	0,298	0,25	25
C1.4	2,44	2,6	2,77	0,052	0,057	0,062	0,127	0,148	0,172	0,15	15
Sum	15,93	17,35	18,93								

As for the criterion C1 that covers the scope of functionality of the ERP system, it should be mentioned that the sub-criterion C1.1 (Accounting and Financial Management) occupies the first place with a weight of 31%. It is important to point out that the highest importance is attached to accounting and financial management functions because accounting is the basic component of information systems of SMEs.

Sub-criterion C1.2 (Purchasing and Sales Management) holds the second position in terms of weight with 29%. This indicates the significance of purchasing and sales management as they affect the efficiency of operations and customer-supplier relationships.

Following this, the sub-criteria C1.3 (Inventory and Logistics Management) comprises 25% weightage. The reason behind assigning a weightage of 25% to inventory and logistics management lies in its significance in ensuring effective physical flow while being less crucial compared to finance and sales functions.

Lastly, sub-criterion C1.4 (Production and Project Management) has the least weight with 15% importance. The reason behind this is that in relation to the SMEs involved in the study, production/project management might not be an urgent need or could be simpler compared to other processes (see Table VII).

2) System Usability: During the decision-making process, where the criterion 'System Usability' will be assessed through the fuzzy AHP method, the following sub-criteria codes will be applied: C2.1, C2.2, C2.3, and C2.4 (see Table VIII). Below are listed the sub-criteria codes together with the sub-criteria themselves:

C2.1: Intuitive user interface

C2.2: Ease of learning the system

C2.3: Ease of navigation

C2.4: Availability of documentation and user support

TABLE VIII. FUZZY PAIRWISE COMPARISON MATRIX OF SUB-CRITERIA FOR THE SYSTEM USABILITY CRITERION

	C2.1			C2.2			C2.3			C2.4		
	l	m	u	l	m	u	l	m	u	l	m	u
C2.1	1	1	1	1,05	1,15	1,30	1,15	1,30	1,45	1,55	1,70	1,90
C2.2	0,77	0,87	0,95	1	1	1	1,05	1,15	1,30	1,45	1,60	1,80
C2.3	0,69	0,77	0,87	0,77	0,87	0,95	1	1	1	1,20	1,35	1,50
C2.4	0,53	0,59	0,65	0,56	0,63	0,69	0,67	0,74	0,83	1	1	1

TABLE IX. FUZZY WEIGHTS AND NORMALIZED PRIORITIES OF SYSTEM USABILITY SUB-CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C2.1	4,75	5,15	5,65	0,054	0,06	0,065	0,257	0,309	0,367	0,31	31
C2.2	4,27	4,62	5,05	0,054	0,06	0,065	0,231	0,277	0,328	0,28	28
C2.3	3,66	3,99	4,32	0,054	0,06	0,065	0,198	0,239	0,281	0,24	24
C2.4	2,76	2,96	3,17	0,054	0,06	0,065	0,149	0,178	0,206	0,17	17
Sum	15,44	16,72	18,19								

In regard to the criterion C2 concerning the system usability, it is evident that the highest weight belongs to the criterion C2.1 (Intuitive user interface). Such a conclusion implies the relevance of the intuitive interface that will allow users to get acquainted with the system and work with it more easily and quickly.

Sub-criteria C2.2 (Ease of learning the system) has the second highest weight of 28%. This highlights that the system's ease of learning is an important criterion, especially within the framework of SMEs, whereby training budgets might be inadequate. It helps that if the ERP system is easier to learn, users can get proficient faster.

Following this, sub-criterion C2.3 (ease of navigation) makes up 24%. It is evident that smooth navigation in the system is significant because it affects the ease with which users can conduct their activities.

Finally, sub-criterion C2.4 (Availability of documentation and user support) is assigned the least weight at 17%, which

can be attributed to the preference by users for a system that is intuitive and easy to operate, thereby minimizing their need for documentation or assistance (see Table IX).

3) Total Cost of the ERP System: During the decision-making process, where the criterion 'Total Cost of the ERP System' will be assessed through the fuzzy AHP method, the following subcriteria codes will be applied: C3.1, C3.2, C3.3, and C3.4 (see Table X). Below are listed the subcriteria codes together with the subcriteria themselves:

C3.1: Cost of ERP licenses

C3.2: Implementation cost

C3.3: Annual maintenance cost

C3.4: User training cost

Regarding criterion C3, dealing with the total cost of the ERP system, the findings reveal that sub-criterion C3.1 (Cost of ERP licenses) occupies the first place with a weight of 32%. The significance attached to the acquisition cost of the software is brought into light from this finding since cost is the major consideration for any SME when choosing an ERP system.

TABLE X. FUZZY PAIRWISE COMPARISON MATRIX OF SUB-CRITERIA FOR THE TOTAL COST OF THE ERP SYSTEM CRITERION

	C3.1			C3.2			C3.3			C3.4		
	l	m	u	l	m	u	l	m	u	l	m	u
C3.1	1	1	1	1,05	1,15	1,3	1,25	1,4	1,55	1,7	1,9	2,1
C3.2	0,77	0,87	0,95	1	1	1	1,15	1,3	1,45	1,6	1,8	2
C3.3	0,65	0,71	0,8	0,69	0,77	0,87	1	1	1	1,15	1,3	1,45
C3.4	0,48	0,53	0,59	0,5	0,56	0,63	0,69	0,77	0,87	1	1	1

TABLE XI. FUZZY WEIGHTS AND NORMALIZED PRIORITIES OF TOTAL COST OF THE ERP SYSTEM SUB-CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C3.1	5	5,45	5,95	0,054	0,059	0,064	0,27	0,322	0,381	0,32	32
C3.2	4,52	4,97	5,4	0,054	0,059	0,064	0,244	0,293	0,346	0,28	28
C3.3	3,49	3,78	4,12	0,054	0,059	0,064	0,188	0,223	0,264	0,23	23
C3.4	2,67	2,86	3,09	0,054	0,059	0,064	0,144	0,169	0,198	0,17	17
Sum	15,68	17,06	18,56								

Sub-criterion C3.2 (Implementation Cost) is the second most important with a weighting factor of 28%. It clearly indicates that the costs related to the implementation process are equally important considerations in selecting an ERP system.

Subsequently, sub-criterion C3.3 (Annual maintenance cost) takes up 23%. This proves that recurring costs are important but usually taken into consideration after initial costs during decision-making.

Lastly, the criterion C3.4 (User training cost) gets the smallest weight, which is 17%. It is because the cost of training users does not seem significant, especially when the system is viewed as an easy-to-use system (see Table XI).

4) Compatibility and Technical Integration: During the decision-making process, where the criterion 'Compatibility and Technical Integration' will be assessed through the fuzzy AHP method, the following subcriteria codes will be applied: C4.1, C4.2, C4.3, and C4.4 (see Table XII). Below are listed the subcriteria codes together with the subcriteria themselves:

C4.1: Compatibility with existing systems

C4.2: Ease of data migration

C4.3: Integration with other software

C4.4: Compatibility with hardware infrastructure

In terms of criterion C4 relating to compatibility and integration, the findings reveal that sub-criterion C4.1 (Compatibility with existing systems) comes first with a weightage of 30%. It indicates the significance of ERP compatibility with the existing systems in the enterprise. High compatibility would help to avoid interruptions, cut down costs, and ensure operational continuity.

TABLE XII. FUZZY PAIRWISE COMPARISON MATRIX OF SUB-CRITERIA FOR THE COMPATIBILITY AND TECHNICAL INTEGRATION CRITERION

	C4.1			C4.2			C4.3			C4.4		
	l	m	u	l	m	u	l	m	u	l	m	u
C4.1	1	1	1	1,05	1,15	1,3	1,1	1,25	1,4	1,45	1,6	1,8
C4.2	0,77	0,87	0,95	1	1	1	1,05	1,15	1,3	1,35	1,5	1,7
C4.3	0,71	0,8	0,91	0,77	0,87	0,95	1	1	1	1,2	1,35	1,5
C4.4	0,56	0,63	0,69	0,59	0,67	0,74	0,67	0,74	0,83	1	1	1

TABLE XIII. FUZZY WEIGHTS AND NORMALIZED PRIORITIES OF COMPATIBILITY AND TECHNICAL INTEGRATION SUB-CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C4.1	4,6	5	5,5	0,055	0,06	0,065	0,253	0,3	0,358	0,3	30
C4.2	4,17	4,52	4,95	0,055	0,06	0,065	0,229	0,271	0,322	0,27	27
C4.3	3,68	4,02	4,36	0,055	0,06	0,065	0,202	0,241	0,283	0,24	24
C4.4	2,82	3,04	3,26	0,055	0,06	0,065	0,155	0,182	0,212	0,18	18
Sum	15,27	16,58	18,07								

Sub-criterion C4.2 (Ease of data migration) comes second with a weight of 27%. The above outcome indicates that one of the difficulties of implementing an ERP solution is related to the ease of data migration.

Thirdly, sub-criterion C4.3 (Integration with other software) takes 24%. From the above outcome, one may understand that the ERP system integration capacity with other software available in the company is also important.

Finally, sub-criterion C4.4 (Compatibility with hardware infrastructure) is assigned the lowest weight – 18%. One explanation could be the adaptability of organizations in the sphere, especially when new technologies are considered (see Table XIII).

5) System Security and Reliability: During the decision-making process, where the criterion 'System Security and Reliability' will be assessed through the fuzzy AHP method, the following subcriteria codes will be applied: C5.1, C5.2, C5.3, and C5.4 (see Table XIV). Below are listed the subcriteria codes together with the subcriteria themselves:

C5.1: Data Protection and Confidentiality

C5.2: User Access Rights Management

C5.3: Data Backup and Restoration

C5.4: System Stability

TABLE XIV. FUZZY PAIRWISE COMPARISON MATRIX OF SUB-CRITERIA FOR THE SYSTEM SECURITY AND RELIABILITY CRITERION

	C5.1			C5.2			C5.3			C5.4		
	l	m	u	l	m	u	l	m	u	l	m	u
C5.1	1	1	1	1,1	1,25	1,4	1,2	1,35	1,5	1,2	1,3	1,45
C5.2	0,71	0,8	0,91	1	1	1	1,05	1,15	1,3	1,05	1,15	1,3
C5.3	0,67	0,74	0,83	0,77	0,87	0,95	1	1	1	1	1,1	1,2
C5.4	0,69	0,77	0,83	0,77	0,87	0,95	0,83	0,91	1	1	1	1

TABLE XV. FUZZY WEIGHTS AND NORMALIZED PRIORITIES OF SYSTEM SECURITY AND RELIABILITY SUB-CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C5.1	4,5	4,9	5,35	0,057	0,061	0,066	0,257	0,299	0,353	0,3	30
C5.2	3,81	4,1	4,51	0,057	0,061	0,066	0,217	0,25	0,298	0,25	25
C5.3	3,44	3,71	3,98	0,057	0,061	0,066	0,196	0,226	0,263	0,23	23
C5.4	3,29	3,55	3,78	0,057	0,061	0,066	0,188	0,217	0,249	0,22	22
Sum	15,04	16,26	17,62								

As for the fifth criterion relating to the system security and reliability, the findings indicate that sub-criterion C5.1 (Data Protection and Confidentiality) has received the highest weight of 30%. This indicates the significance attached to the protection of confidential data and information confidentiality given that the target organizations deal with sensitive information.

The sub-criterion C5.2 (User Access Rights Management), coming in second place with the weight of 25%, reflects the importance of the access rights management in terms of information access control and the enhancement of security.

Thirdly, sub-criterion C5.3 (Data Backup and Restoration) has 23% weight, demonstrating that the mechanisms of data restoration are crucial in order to ensure the business continuity in case of incidents.

Finally, the sub-criterion C5.4 (System Stability), which is of low importance, has 22% weight. Despite being significant as well, the issue is less prioritized than others as it is usually considered one of the basics of the ERP system (see Table XV).

6) Scalability of the ERP system: During the decision-making process, where the criterion 'Scalability of the ERP System' will be assessed through the fuzzy AHP method, the following sub-criteria codes will be applied: C6.1, C6.2, C6.3, and C6.4 (see Table XVI). Below are listed the sub-criteria codes together with the sub-criteria themselves:

C6.1: Ability to add new modules

C6.2: Adaptability to business growth

C6.3: Ability to customize the system

C6.4: Availability of future functional enhancements

TABLE XVI. FUZZY PAIRWISE COMPARISON MATRIX OF SUB-CRITERIA FOR THE SCALABILITY OF THE ERP SYSTEM CRITERION

	C6.1			C6.2			C6.3			C6.4		
	l	m	u	l	m	u	l	m	u	l	m	u
C6.1	1	1	1	1,05	1,15	1,3	1,1	1,25	1,4	1,2	1,35	1,5
C6.2	0,77	0,87	0,95	1	1	1	1,05	1,15	1,3	1,1	1,25	1,4
C6.3	0,71	0,8	0,91	0,77	0,87	0,95	1	1	1	1,05	1,15	1,3
C6.4	0,67	0,74	0,83	0,71	0,8	0,91	0,77	0,87	0,95	1	1	1

TABLE XVII. FUZZY WEIGHTS AND NORMALIZED PRIORITIES OF SCALABILITY OF THE ERP SYSTEM SUB-CRITERIA

	Σl	Σm	Σu	1/Σu	1/Σm	1/Σl	S_l	S_m	S_u	W	%
C6.1	4,35	4,75	5,2	0,056	0,061	0,067	0,244	0,29	0,348	0,29	29
C6.2	3,92	4,27	4,65	0,056	0,061	0,067	0,22	0,26	0,312	0,26	26
C6.3	3,53	3,82	4,16	0,056	0,061	0,067	0,198	0,233	0,279	0,24	24
C6.4	3,15	3,41	3,69	0,056	0,061	0,067	0,176	0,208	0,247	0,21	21
Sum	14,95	16,25	17,7								

With regard to criterion C6 concerning the scalability of the ERP system, sub-criterion C6.1 (Ability to add new modules) ranks first and is assigned the highest weighting among all other sub-criteria: 29%. The implication is that it is crucial for the system to develop incrementally and be able to incorporate more modules when required by the organization.

The sub-criterion C6.2 (Adaptability to business growth), with a weightage of 26% of the total evaluation criteria, means the ability of the system to adapt and aid in the growth of the business is an important factor in the current scenario.

Sub-criterion C6.3 (Ability to customize the system) carries 24% weightage. The fact that the customization ability of the system based on the company's requirements is an important criterion but not as critical as scalability-related criteria is evident from the above result.

Last, criterion C6.4 (Availability of future functional enhancements) has the least weight with 21%. This may be attributed to the fact that future improvements are considered less urgent or less important in the early stages of decision-making (see Table XVII).

VII. PROPOSED MODEL FOR SELECTING ERP FOR MOROCCAN SMES

A. General Principle of the Model

As the basis of transformation of findings acquired using the Fuzzy AHP method into something tangible and practically usable by organizations, the study puts forward a decision model that makes possible objective comparison of various ERP solutions available to Moroccan SMEs. It is a hierarchically structured model composed of a decision goal, main selection criteria, and operational sub-criteria, which allow for a realistic assessment of ERP solutions. Main criteria are the general aspects affecting the decision about the adoption of an ERP application, while operational sub-criteria are more tangible and observable functional features of ERP applications.

B. Mathematical Formulation of the Model

Let C_i be a selection criterion, where $i = 1, \ldots, 6$, and SC_{ij} a sub-criterion belonging to criterion C_i, where $j = 1, \ldots, 4$.

Fuzzy AHP analysis allows us to obtain the relative weight of each criterion, denoted w_i, reflecting its importance in the overall decision. Within each criterion, the sub-criteria have local weights denoted v_{ij}, such that:

$$\sum_{j=1}^{4} v_{i,j} = 1$$

To determine the actual weight of each sub-criterion in the final decision, an overall weight is calculated as follows:

$$g_{i,j} = w_i \times v_{i,j}$$

The consistency of the model is ensured by:

$$\sum_{i=1}^{6} \sum_{j=1}^{4} g_{i,j} = 1$$

C. Evaluation of the ERP Solutions

Each candidate ERP solution is evaluated by the company's experts or users using a score between 0 and 10 for each sub-criterion, denoted as $(s_{i,j}^{(p)})$, where (p) represents the ERP solution being analyzed.

The overall score for an ERP solution is calculated as follows:

$$Score\left(ERP^{(p)}\right) = \sum_{i=1}^{6} \sum_{j=1}^{4} g_{i,j} \times s_{i,j}^{(p)}$$

The selected solution is the one that achieved the highest score:

$$ERP^{*} = arg\ max\ Score(ERP)$$

This mechanism thus makes it possible to convert decision-makers' qualitative assessments into a quantitative metric, making it easier to compare ERP solutions.

D. Operational Evaluation Table for ERP

Table XVIII shows the operational approach for comparing ERP solutions based on the proposed model. The rows of the table indicate the sub-criteria that need to be considered for evaluating the ERP. These sub-criteria include aspects like accounting management, sales management, data security, and technical compatibility.

The Weight (g) column indicates the relative importance of each sub-criterion in the final decision, as determined by the Fuzzy AHP analysis. These weights reflect the priorities of Moroccan SMEs when selecting an ERP.

The columns (ERP_1 Note), (ERP_2 Note) allow you to assign notes to each ERP solution on a scale of 0 to 10.

The columns (ERP_1 Score) and (ERP_2 Score) show the weighted scores calculated for each ERP solution by multiplying the Weight (g) by the corresponding note.

The table below thus serves as the primary calculation tool for converting the qualitative assessment of ERP solutions into a comparable quantitative score.

TABLE XVIII. OPERATIONAL EVALUATION TABLE FOR ERP

Code	Sub-criteria	Weight (g)	ERP_1 Note	ERP_2 Note	ERP_1 Score	ERP_2 Score
C1.1	Accounting and Financial Management	0,083				
C1.2	Purchasing and Sales Management	0,078				
C1.3	Inventory and Logistics Management	0,068				
C1.4	Production and Project Management	0,04				
C2.1	Intuitive user interface	0,068				
C2.2	Ease of learning the system	0,061				
C2.3	Ease of navigation	0,053				
C2.4	Availability of documentation and user support	0,039				
C3.1	Cost of ERP licenses	0,058				
C3.2	Implementation cost	0,053				
C3.3	Annual maintenance cost	0,041				
C3.4	User training cost	0,031				
C4.1	Compatibility with existing systems	0,043				
C4.2	Ease of data migration	0,038				
C4.3	Integration with other software	0,034				
C4.4	Compatibility with hardware infrastructure	0,026				
C5.1	Data Protection and Confidentiality	0,033				
C5.2	User Access Rights Management	0,028				
C5.3	Data Backup and Restoration	0,02				
C5.4	System Stability	0,024				
C6.1	Ability to add new modules	0,024				
C6.2	Adaptability to business growth	0,021				
C6.3	Ability to customize the system	0,019				
C6.4	Availability of future functional enhancements	0,017				
	Total score	1				

E. Procedure for Using the Table

1) Assign an ERP Note between 0 and 10 to each ERP solution for each sub-criterion.

2) Calculate the ERP Score for each sub-criterion using the following formula:

$$ERP\ Score = \text{Weight (g)} * ERP\ Note$$

3) Add the weighted scores to obtain the total score for each ERP solution.

4) Compare the scores and select the solution with the highest score.

F. Contribution of the Proposed Model

The model maintains the scientific rigor of multiple criteria decision-making without becoming difficult to use in a business setting. In this way, Moroccan SMEs are ensured an

appropriate methodology for the comparison of ERP systems and reducing the likelihood of a bad choice.

VIII. Discussion

The results obtained through the application of the Fuzzy AHP method reveal a clear hierarchy of selection criteria for ERP systems in the context of Moroccan SMEs. The analysis reveals that criteria related to the system's functional scope rank first, followed by ease of use and the overall cost of the system. This prioritization reflects a marked preference among SMEs for solutions that are immediately operational, capable of effectively meeting business needs while remaining simple to use.

These results are consistent with several studies in the literature. For example, [3] demonstrated that critical factors related to ERP selection, particularly system performance and integration, play a decisive role in the decision-making process. Similarly, [8] emphasize the importance of hybrid multi-criteria approaches for improving the ranking of ERP solutions by taking multiple dimensions into account.

The emphasis placed on ease of use in our study which confirms that user adoption is a key factor for success. This finding aligns with the conclusions of [9], which highlight the central role of system quality, particularly in terms of usability and reliability. Similarly, [6] shows that incorporating multiple criteria, including user-related aspects, improves the relevance of decisions.

Furthermore, the total cost of the system emerges as a determining criterion in our study, reflecting the budget constraints specific to SMEs. This finding is also confirmed by [4], who identify cost as a key factor in the evaluation of ERP solutions, particularly in industrial environments.

In contrast, criteria related to scalability and future developments appear to be of lower priority in our study. This finding may be interpreted as reflecting SMEs' focus on immediate needs rather than long-term prospects. This trend differs in part from certain studies, such as [11], which place greater emphasis on long-term strategic and organizational factors in ERP selection.

Finally, the model proposed in this study serves as a relevant decision-making tool, helping to structure the ERP selection process and reduce the subjectivity of choices. This approach builds on the work of [10], who highlight the value of multi-criteria methods for improving the quality of decisions in complex contexts.

IX. Conclusion

The objective of this study was to identify and prioritize the selection criteria for ERP systems in the context of Moroccan SMEs, as well as to propose a decision-making model tailored to their needs. To achieve this, an approach combining qualitative interviews and the Fuzzy AHP method was adopted, allowing for both the experience of professionals and a structured evaluation of the criteria to be taken into account.

This research aimed at defining and prioritizing selection criteria for ERP systems within the context of SMEs in Morocco and developing a decision model based on the needs of such companies. In order to reach these objectives, the choice was made to use a mixed methodology, including interviews and application of the fuzzy AHP method, which takes into consideration both professional experience and criteria evaluation.

The results highlight the top priority of the ERP system's functional scope, followed by ease of use and the overall cost of the system. These results reflect a clear preference among SMEs for practical, accessible, and immediately operational solutions. Furthermore, technical criteria, such as compatibility and integration, as well as system security, remain important but are considered secondary to immediate functional needs. Finally, system scalability appears to be the lowest priority, suggesting that SMEs prioritize solutions that meet their current needs rather than a long-term vision.

A detailed analysis of the sub-criteria confirms this trend, highlighting the importance of financial management features, system usability, and initial costs. These factors reflect the specific constraints faced by SMEs, particularly in terms of limited resources, the need for simplicity, and the need for rapid implementation.

Based on these results, a decision-support model has been proposed to objectively evaluate and compare different ERP solutions by weighting the criteria and sub-criteria. This model transforms qualitative judgments into a quantitative assessment, thereby facilitating the selection process. It serves as an operational tool accessible to SMEs, enabling them to reduce the risks associated with choosing an unsuitable system and to improve the quality of their decision-making.

In conclusion, this research makes both a theoretical and practical contribution. On the one hand, it enriches existing work on ERP selection by incorporating the specific context of Moroccan SMEs. On the other hand, it offers a practical tool that companies can use to structure their selection process. However, certain limitations should be noted, particularly the sample size and the contextual nature of the results. Future research could expand the study to include a larger number of companies or incorporate other multi-criteria methods to strengthen the robustness of the proposed model.

References

[1] Cruz-Cunha MM, Silva JP, Gonçalves JJ, Fernandes JA, Ávila PS (2021) ERP selection using an AHP-based decision support sys tem. In: Research anthology on recent trends, tools, and implica tions of computer programming, IGI Global, Pennsylvania, pp 374–391.

[2] Sánchez, P. J., Martı, L., Garcı, C., Herrera, F., & Herrera-Viedma, E. (2009). A fuzzy model to evaluate the suitability of installing an enterprise resource planning system. Information Sciences, 179(14), 2333-2341.

[3] Alherimi, N., Alyaarbi, A., Ali, S., Bahroun, Z., & Ahmed, V. (2025). Prioritizing ERP System Selection Challenges in UAE Ports: A Fuzzy Delphi and Relative Importance Index Approach. Logistics, 9(3), 98.

[4] Dağci Yüksel, B., & Ersöz, F. (2024). Evaluation of ERP software selection criteria with fuzzy AHP approach: an application in the metal production enterprises in the aviation industry. International Journal of System Assurance Engineering and Management, 15(6), 2656-2667.

[5] Bhatt, N., Guru, S., Thanki, S., & Sood, G. (2021). Analysing the factors affecting the selection of ERP package: a fuzzy AHP approach. Information Systems and e-Business Management, 19(2), 641-682.

[6] Ayağ, Z., & Yücekaya, A. (2019). A fuzzy ANP-based GRA approach to evaluate ERP packages. International Journal of Enterprise Information Systems (IJEIS), 15(1), 45-68.

[7] Haddara M, Zach O (2011) ERP systems in SMEs: a literature review. 2011 44th Hawaii International Conference on System Sciences, 1–10.

[8] Akpinar, I., Kececi, B., Atalay, K., & Boran, M. (2025). ERP software selection problem with a novel Hybrid hesitant fuzzy cross-entropy and probabilistic hesitant fuzzy MCDM approach. Journal of Industrial and Management Optimization, 21(9), 5662-5686.

[9] Utama, D. M., Ibrahim, M. F., & Jabari, A. N. A. A. K. (2024). A fuzzy multi-criteria approach for selecting open-source ERP systems in SMEs using fuzzy AHP and TOPSIS. Jurnal Optimasi Sistem Industri, 23(2), 167-187.

[10] Pereira, T., & Ferreira, F. A. (2020). An ERP selection using a hybrid methodology. Journal of Computational Methods in Science and Engineering, 20(3), 689-695.

[11] Gupta, R., & Naqvi, S. K. (2017). A framework for applying CSFs to ERP software selection: an extension of fuzzy TOPSIS approach. International Journal of Intelligent Information Technologies (IJIIT), 13(2), 41-62.

[12] Saaty, T. L. (1980). The analytic hierarchy process.

[13] Chang, D. Y. (1996). Applications of the extent analysis method on fuzzy AHP. European journal of operational research, 95(3), 649-655.